2018

The Feasibility of Dementia Caregiver Task Performance Measurement Using Smart Gaming Technology

Garrett G. Goodman
Wright State University

The Feasibility of Dementia Caregiver Task Performance Measurement Using Smart Gaming Technology

A Thesis submitted in partial fulfillment
of the requirements for the degree of
Master of Science

by

Garrett G. Goodman
B.S.C.S., Wright State University, 2016

2018
Wright State University

Wright State University
GRADUATE SCHOOL

November 7, 2018

I HEREBY RECOMMEND THAT THE THESIS PREPARED UNDER MY SUPERVISION BY Garrett G. Goodman ENTITLED The Feasibility of Dementia Caregiver Task Performance Measurement Using Smart Gaming Technology BE ACCEPTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF Master of Science.

Tanvi Banerjee, Ph.D.
Thesis Director

Mateen Rizki, Ph.D.
Chair, Department of
Computer Science and Engineering

Committee on
Final Examination

Tanvi Banerjee, Ph.D.

William Romine, Ph.D.

Amit Sheth, Ph.D.

Jennifer Hughes, Ph.D.

Barry Milligan, Ph.D.
Interim Dean of the Graduate School

ABSTRACT

Goodman, Garrett G. M.S., Department of Computer Science and Engineering, Wright State University, 2018. *The Feasibility of Dementia Caregiver Task Performance Measurement Using Smart Gaming Technology*

Dementia caregiver burnout is detrimental to both the familial caregiver and their loved ones with dementia. As the population of older adults increases, both the number of individuals with dementia and their corresponding caregivers increase as well. Thus, we are interested in developing a potential tool to non-invasively detect signs of caregiver burnout using a mobile application combined with machine learning. Hence, the mobile application "Caregiver Assessment using Smart Technology" (CAST) was developed which personalizes a word scramble game. The CAST application utilizes a heuristically constructed Fuzzy Inference System (FIS) optimized via a Genetic Algorithm (GA) to provide an individualized performance measure for each user of CAST. That is, we attempt to adjust the difficulty of the game using an individual user's ability to solve the word scramble tasks. With a cohort of 48 non-caregiver participants and 2 dementia caregiver participants, we report on the construction of the FIS, the optimization of the FIS using the GA, and analysis of the preliminary results of deciding difficulties of words using standard performance metrics precision, recall, and F1 score.

Contents

List of Figures

List of Tables

Acknowledgments

To Dr. Tanvi Banerjee, thank you for giving me the opportunity to learn under you and to be a part of your team. You took me under your wing as a student without knowing any of my skills or knowledge, only knowing that I had a desire to teach. Before becoming your student, I was afraid that I would a programmer for my entire life with no way of pursuing my dream.

I would like to thank Dr. Jennifer Hughes for supporting me with my writing and for providing fun and interesting challenges through my tenure with the research group. She is always very supportive and put me in situations where I learned more than I did in some classes. Thank you to JoAnna, Morgan, Abby, and Alex for being great team members, always making me laugh on the job, and taking me out of my comfort zone as a computer scientist. Finally, I would like to thank Cogan and Iosif for proof reading my work and Quinn for help in improving the quality of my figures.

This thesis was supported in part by the NIH under grant K01 LM012439-01. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the National Institute of Health.

This thesis is dedicated to my father, Major Eric Gwyn Goodman, USAF, Retired.

1

Introduction

Degenerative neurological diseases, such as Alzheimer's disease, are becoming more common due to the rise in the older population [1]. Alzheimer's disease is the most prevalent disease under the dementia umbrella [1]. People with this disease experience increased neurological and functional decline as the disease progresses through its early, middle, and late stages [6]. Specifically, Alzheimer's is a chronic neurodegenerative disease that destroys neural connections in the brain [14]. This results in memory impairment, speech deficiency, decline in motor skills, and more, [6] which inevitably leaves the individual vulnerable and unable to be independent. Furthermore, Alzheimer's disease currently affects 1 in 10 Americans aged 65 or older which equates to approximately 5.7 million Americans [1]. The projected amount of older adults that could be affected by Alzheimer's disease and other forms of dementia such as Huntington's disease, Parkinson's disease, Vascular dementia, etc. [2] is estimated at 14 million by 2050 [1]. Due to the diseases deterioration of the brain, Alzheimer's disease is the sixth leading cause of death in Americans [1].

Managing Alzheimer's disease and other forms of dementia is difficult, expensive, and largely performed by a family member such as a spouse or child [1]. Hence, an estimated 16.1 million Americans provide unpaid care to family members affected by Alzheimer's disease and other forms of dementia, resulting in 83% of the total care [1]. Though, ad-

ditional paid professional care is also sought after and costs more than 17% of the United states gross domestic product [22, 34]. With the time and challenges that come with caring for a family member with Alzheimer's disease, it is understandable that a large amount of stress is also associated with caregiving. Dementia caregivers must assist with many things such as activities of daily living (ADLs) which include activities such as bathing, eating, dressing, and toileting [12, 30]; and the instrumental activities of daily living (IADLs) which include medication management, transportation, housework, and emotional support [12, 30]. Consequently, providing care causes stress and affects the quality of life of caregivers [8, 13]. Although being the primary caregiver for a family member with dementia is stressful, it is beneficial to both the caregiver and society as it allows the individual with the disease to remain in their home for longer. The symptoms of this overwhelming stress consists of emotional exhaustion, depersonalization, and reduced sense of personal accomplishment [38]. Also, research has shown a correlation between caregiver health decline and stress [26], leading to caregiver burnout.

With this information, it is important to investigate possibilities to assist the family caregiver. In this thesis, we focus on incorporating state of the art and unobtrusive technologies to be used as the underlying structure in the identification of caregiver stress. An early detection of high stress in a dementia caregiver can be useful in early detection of caregiver burnout. Furthermore, a successful detection can be used as a factor in determining when a caregiver and their corresponding family member with dementia could use external assistance via community intervention. Additional caregiver support through community and unobtrusive automatic detection systems can reduce stress and possibly prolong or prevent the institutionalization of the family member with dementia.

Research has shown that familial dementia caregivers show interest in such technologies which can assist them in their caregiving. The American Association of Retired Persons Project Catalyst & Healthcare Innovation Technology Laboratory 2016 survey [28] shows that Alzheimer's caregivers show interest in technologies which can assist them in

their role as a caregiver. Furthermore, Burstein et al. concluded that caregivers are interested in technologies to assist their role in caregiving [7] by examining caregivers' knowledge and skill of existing technology. Following, Maiden et al. focused their research on creating a digital life history manager application for mobile devices to assist caregivers in their duties [27]. Specifically, the app stored patient information, social contact with other caregivers, and patient history [27]. This app was proposed to aid the caregivers via new technologies and to introduce digital data into the caregiving realm. Finally, research from Hughes et al. determined that caregivers are interested in unobtrusive technology that could identify stress [20]. This study looked at a preliminary mobile application called the Caregiver Assessment using Smart Technology (CAST) application [20]. The CAST app proposes using a simple game on a mobile device (e.g. smartphone or tablet) as a tool to monitor task performance of the dementia caregivers, temporally [3, 20].

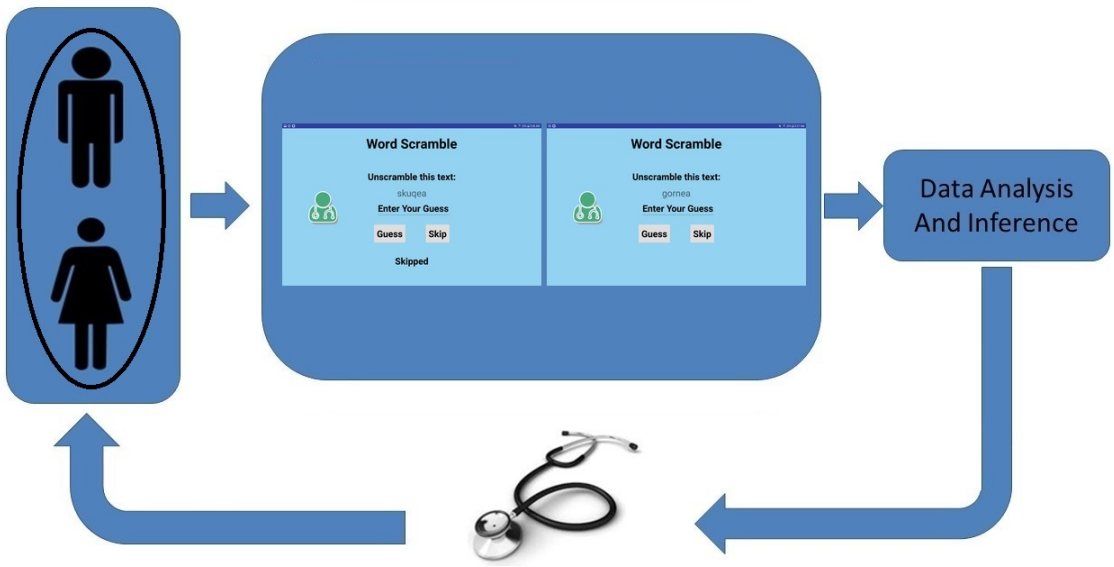

Figure 1.1: The prospective CAST operation loop begins on the left with the dementia caregiver (circled). The caregiver plays the proposed word scramble game on the CAST app which saves information about the user interaction and gameplay. Using the saved information, it will check for any indicators of anomalous stress and will forward the information to a clinician of choice to discuss community intervention.

By using the CAST app described in the study by Hughes et al. [20], we focus this thesis on the creation and evaluation of a Fuzzy Inference System (FIS). This FIS will be used as the underlying system for the task performance detector in the CAST word scramble game. The word scramble game was chosen as it is inherently task oriented and therefore susceptible to discretization. As each word is to be considered a task, we can create Individualized Word Difficulties (IWD) which can adapt based upon user interaction and gameplay features, as discussed in Section 3.6. As the name suggests, the IWD differs from person to person which allows for individualized detection even for multiple people using the same device. By creating three difficulties of Easy, Medium, and Hard for the IWD, users can progress and regress through the three difficulties. This allows the system to represent changes in the users ability levels. Therefore, having a system capable of making objective classifications of word difficulties is important as it is the baseline before classifying task performance and possibly stress. This overall goal is represented in Figure 1.1, where the caregiver (circled) plays the word scramble game on the CAST app. Following, the data is sent through the proposed FIS. Finally, if any anomalies are found, the information is sent to a clinician of choice for them to discuss options to the dementia caregiver, such as community intervention.

The method used in this thesis differs from the normal means of measuring stress as we take an indirect approach. That is, traditional methods of measuring stress involve the use of perceived stress scales [9] in the form of self conducted surveys; or invasive cortisol measurements from blood, urine, or saliva [24]. Our method means to first establish a system of making objective classifications of a word scramble game (this thesis), then monitoring the classifications for attempting to make task performance predictions and following, potential stress predictions. As this thesis is focused on creating and evaluating the FIS for the CAST app word scramble game, we address the following research questions about the proposed FIS creation and evaluation:

RQ1. How can we build an FIS system that incorporates user and gameplay features in

measuring IWD?

RQ2. Are certain words better at discriminating between different participants ability levels?

RQ3. Using established performance metrics, how does the FIS system constructed from RQ1 perform?

Chapter Overview

The remainder of this thesis is organized as follows.

Chapter 2: Related Work discusses the research done on caregiver stress assessment, gaming as a diagnostics tool, feature extraction from gaming, the links from task performance to stress, and how gaming can be used to detect task performance.

Chapter 3: Methods details the dataset used for this thesis, the data preprocessing steps using the Rasch Model, the construction of the FIS, and the implementation of the GA to optimize the FIS.

Chapter 4: Results examines the results of the preprocessed data with both the heuristically created FIS and the optimized FIS using the GA.

Chapter 5: Discussion provides analysis of the chosen word set used in the word scramble game and the results for both the heuristically created FIS and and GA optimized version.

Chapter 6: Conclusion summarizes the contributions from this thesis and restates the three research questions with their corresponding answers.

Chapter 7: Future Work discusses potential future work that can be done to improve the results of the study and also expand on the overall concepts.

2

Related Work

In this chapter, we discuss the studies related to the approach taken in this thesis. Specifically, five areas are examined to make the connection of using gaming and machine learning to measure task performance and are organized as follows. Section 2.1 describes studies addressing caregiver stress assessment and its importance. Following, Section 2.2 discusses the use of gaming for clinical diagnostics and for medical assistance. Next, Section 2.3 explores the ability to extract features for a machine learning algorithm via gaming. Section 2.4 outlines research on how task performance can be measured with gaming. Finally, Section 2.5 summarizes studies that utilize task performance as a biomarker for stress and cognitive performance.

2.1 Caregiver Stress Assessment

It is important to acknowledge the overwhelming stress experienced by the familial caregiver of an individual with dementia. Acknowledgment can allow for stress management to improve both the physical and mental health of the caregiver and thus, the patient with a dementia related disease as well. A study by Lu and Wykle used a sample size of 99 caregivers of people with dementia in conjunction with a cross-sectional correlation de-

sign for determining the effects of caregiver stress on health and functionality [26]. Their findings show a relationship between caregiver stress, functional ability, and self-care measures. Furthermore, the caregivers that reported high levels of stress also had poorer health, depression, and lower functional ability. Following, research also shows a correlation of caregiver stress and negative patient outcomes [5, 36]. With respect to familial caregivers of dementia patients, the most common and least favorable outcome is the institutionalization of the patient. These studies further show the need of reliable caregiver stress assessment and management.

2.2 Gaming as a Diagnostics Tool

As games are more frequently seen in today's society, they become a more viable option to be used for clinical diagnostics or some form of medical assistance. For gaming as a diagnostic tool, Ranjbartabar et al. described a framework required for games to be used for clinical diagnostics [31]. Their study measured stress levels using the linguistic variables very low, low, high, and very high. The authors discuss that such a gaming application could be used as a clinical diagnostic tool for clinical issues such as Post Traumatic Stress Disorder (PTSD), exam stress, depression, acute stress disorder, and more [31]. Another study by Dillon et al. created a gaming application and found statistically significant differences of physiological measures in participants during gameplay with induced stressors [10]. Furthermore, researchers at the Institute for Research in Computer Science and Automation created a game using the digit-symbol substitution method for memory assessment [37]. This game came equipped with an animated virtual conversational agent (virtual avatar modeled after a human) to mimic a therapist, which produced similar results to using a real therapist for the task. The presented studies show the plausibility of using gaming as a way to measure stress or in other diagnostic instances.

2.3 Feature Extraction Using Gaming

Video games are generally complex with a multitude of things happening at once, even for the most basic of games. This allows for the possibility of recording the actions taken during gameplay and the corresponding effects to be used as features in a machine learning algorithm. Research from Harpstead et al. created analysis tools to capture and analyze performance of students within educational games [18]. They used heuristically created features from the educational game RumbleBlocks, a game that students must build a structurally sound tower while completing side objectives [18]. The three features were an empirical measure of symmetry, width of the base of the tower, and the center of mass of the tower. The authors proceeded to implement logistic regression to predict level success to measure student learning within the game.

Similarly, Harpstead et al. did a second study on the same subject where conceptual features were extracted from the same RumbleBlocks game [17]. A four step process is used that ingests all of the log files of built towers from a student, generates grammatical rules of the towers using their own unsupervised method called the Exhaustive Rule Generator, parses the grammar results into hierarchical trees, then converts them into numerical features [17]. The final feature vectors represent the grammatical concepts represented from the rules of the towers, which are now compatible for a machine learning algorithm.

Finally, Southey et al. created an active learning tool for automated analysis of the soccer video game FIFA '99 for assisting game designers in achieving predetermined gameplay goals [33]. These goals can be anything from estimated game completion time to perceived game difficulty. The authors determined that the most important metrics are directly related to the gameplay, such as time elapsed, goals scored, and more [33]. This approach is similar to our approach of feature extraction from gameplay further discussed in Section 3.6.1. The three presented studies show the different approaches to feature extraction from gameplay and that it is both plausible and case dependent.

2.4 Gaming in Relation to Task Performance

The ability to utilize a game to measure task performance is a necessary area of study for both stress and other medical issues such as PTSD. Research from Holmgard et al. utilized a virtual world in which a player is to shop at a supermarket [19]. The player completes tasks in the environment that are designed to elicit stressors for treatment and intervention of patients suffering from PTSD. The game tracks the performance of the player over a week-long period in which researchers were able to determine the existence of PTSD symptoms and at what times they occurred [19]. Although this study was preliminary in nature, it provides the necessary support for the claim that gaming can be used as a task performance measuring platform. It follows that the same claim can be theoretically supported for primary caregivers of dementia patients as well.

2.5 Task Performance as a Biomarker

Biomarkers, a measurable characteristic that is indicative of normal medical responses, biological processes, pathogenic process, and pharmacological responses [11], are imperative for predicting outcomes of diseases [35], and in our case, stress. Research has been done that utilizes task performance as a biomarker. A study from Gutshall et al. examined different types of stressors effects on working memory, problem solving, and decision making by using task performance to measure changes in stress level [15]. The study was performed over a two-week period with police officers of varying experience performing duties with low, moderate, and acute stress. Police officers experiencing acute stress showed significant impairments in their working memory [15]. This result shows that high stress levels, similar to those which cause caregiver burnout, have the ability to debilitate the task performance of caregivers.

Stress has also been shown to affect cognitive functionality as well. Research per-

formed by Korten et al. used a cohort of 1099 older adults between 64 and 100 years of age [23]. The study measured the participants stress levels and cognitive functionality. The results showed that participants which had higher stress levels performed poorer on task related problems such as ordering and backward digit span than the less stressed counterparts [23]. This result depicts the plausibility of a game being used as a task performance monitoring system for stress.

These five subsections of research show the link from gaming to using task performance as a biomarker for stress. This thesis specifically discusses the means of creating and evaluating a system that is capable of making task performance predictions. Following, the presented research neglects using task performance or stress evaluation in the caregiver realm. Furthermore, the research discusses immediate intervention rather than providing or creating a baseline for consistent evaluation. Due to the inherent individuality of caregivers and the range of the caregiver burden spectrum, it is important to create an adapting system that takes into account individuals.

Our research differs from the related research as it incorporates task performance measurement via the proposed word scramble game of the CAST app and an underlying FIS. This proposed method takes an indirect approach to task performance measurement, and potentially stress, as it is unobtrusive and differs from the standard stress measurement techniques (e.g. cortisol measurement [24] and subjective personal stress scales [9]). Finally, our proposed method provides information about why individual participants achieved a specific difficulty for a word. This is due to the inherent explainability of FIS [21, 29], which is a useful factor in the medical realm.

3

Methods

In this chapter, we discuss the methods used to collect and process our dataset. Specifically, Section 3.1 gives an outline of the CAST application's word scramble game. Following, Section 3.2 describes the data collection procedure. Next, Section 3.3 describes the collected dataset in detail. The dataset preprocessing steps are deliberated on in Section 3.4 for being used in the FIS. Then, Section 3.5 will discuss the qualitative survey which was created to test the intuitiveness of the applications design and usage. Section 3.6 presents the framework of the heuristically created FIS. Finally, Section 3.7 discusses the GA used to optimize the FIS.

3.1 Application Description

This thesis utilizes the CAST application's word scramble game to gather our data. Figure 3.1 shows the word scramble game in progress on a Samsung Galaxy Tab A [32] Android tablet. The app was designed to purposely have a simplistic and intuitive layout. Hence, the only three interactive options on the screen are the "Guess" and "Skip" buttons as well as the "Enter Your Guess" text field. When the word scramble game begins, the user is presented with a string of letters which unscrambles to a specific word. The example in

Figure 3.1 shows the string 'hcekc' which unscrambles to 'check'. The user would proceed to press the "Enter Your Guess" text field so that the tablets keyboard would appear for them to enter their guess. Once a guess is finalized, the user would press the "Guess" button where two outcomes can occur. Either the entered guess was incorrect and the user would be notified by the text field being cleared and by displaying the word "Incorrect" at the bottom of the screen. The user would then attempt another guess. If correct, a pop-up box will appear asking the user to rate the difficulty of the task from 1-10 where 1 is easy and 10 is hard, as shown in Figure 3.2. This User Rated Difficulty (URD) will be one of two ground truths to be used with the FIS. The game would then proceed and present a new string of letters for the user to solve until the game is finished. Note that this pop-up rating system is only for the validation for the algorithm and is intended to be removed in the future. Finally, the last option is for the user to press the "Skip" button that would present the same pop-up rating box for the user, then continue with a new word task. There is no upper bound on how many guesses a user can attempt nor is there a time limit.

The CAST word scramble game was built using Android Studio version 3.1.0 and operates with Android version 4.3 or above. The default language for Android Application Programming Interface 27 is used and no external packages are included. During gameplay, data is saved on a per word basis directly on the device. That is, a SQLite database was designed for the application data storage. Version 3.19 of the SQLite database was used in the CAST app. There are no external wireless connections to the database. Thus, to retrieve information from the database, a manual connection to a computer is required.

3.2 Data Collection

Institutional Review Board (IRB) approved data collection for this thesis occurred from two groups and additional individual recruitment while using the CAST application's word

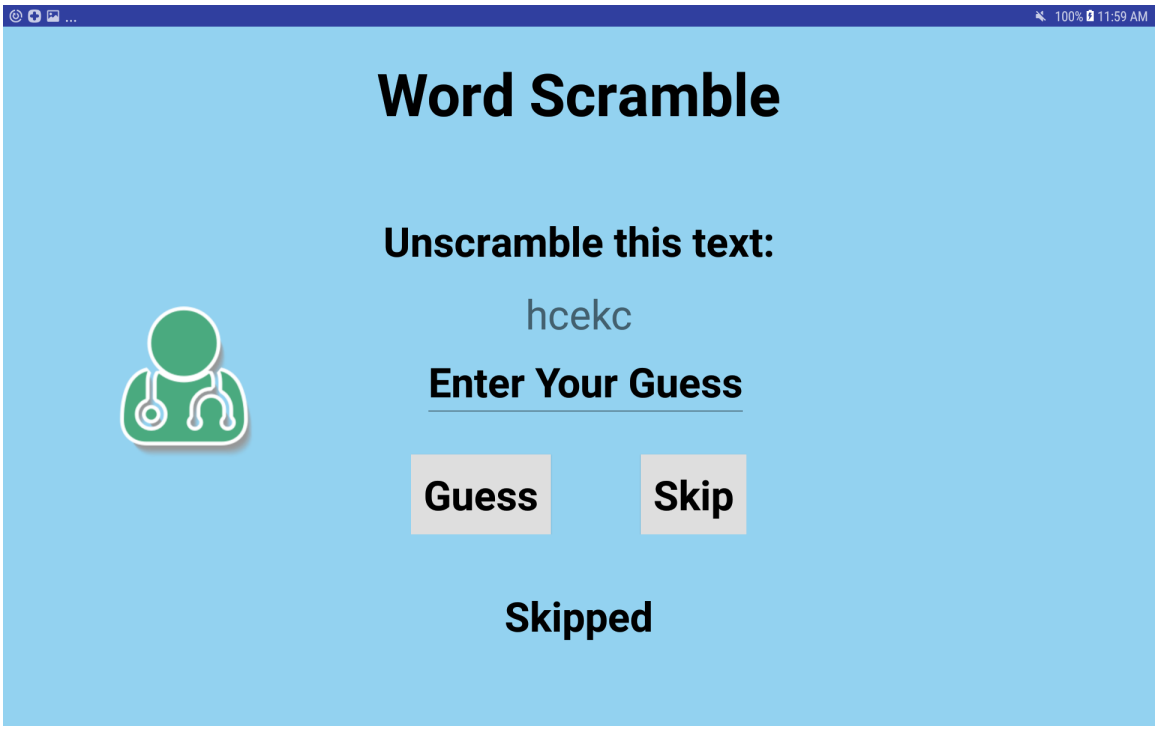

Figure 3.1: The CAST application's word scramble game in progress. The user interface, both simplistic and intuitive, only has three interactive items. The user would examine the string of letters, in this case 'hcekc', and press the "Enter Your Guess" text field to make a guess. Once the guess is inputted, the user would press the "Guess" button to guess. If the user is unable to complete the task, they would proceed to press the "Skip" button to skip the current word.

scramble game. The two groups were the Zeta Tau Alpha (Eta Pi chapter) sorority and a class of graduate-level social work students from Wright State University. The individually recruited participants came from research colleagues and their corresponding professors. The final cohort consists of 48 participants. The participants ages range from 20-60 years old and all participants at minimum were pursuing a baccalaureate degree. Each participant played a single game of the word scramble which consisted of 28 words each, resulting in 1,344 data points. Of the gathered data points, 24 were removed due to the participants failing to provide a difficulty rating for the word in question, resulting in a final dataset of 1320 data points. During gameplay, the order of the words were consistent among all participants to prevent presentation bias. The correct answers were also presented after the

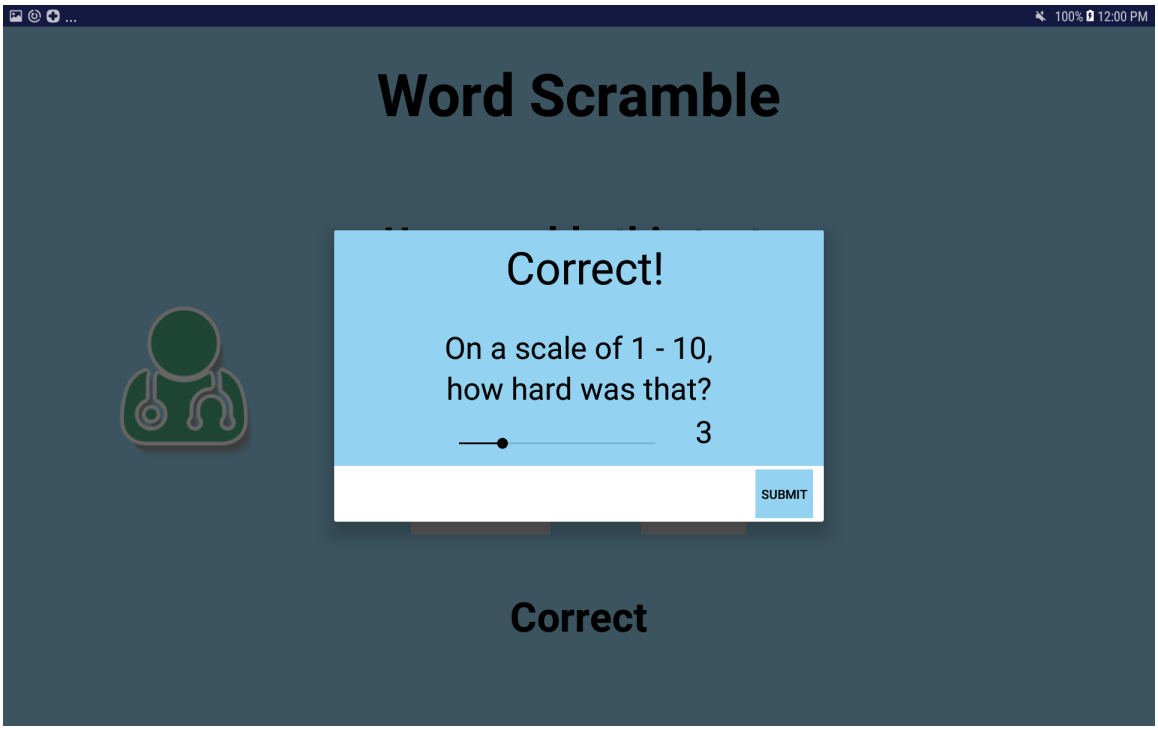

Figure 3.2: The pop-up box after either correctly unscrambling a word or skipping the task. The text "On a scale of 1-10, how hard was that?" is displayed where 1 is Easy and 10 is Hard. To input the URD, a slide bar is given where the integer on the right of the slide bar is updated in real time, with respect to the position of the slide bar. Finally, a "SUBMIT" button is given the bottom right hand corner of the pop-up box to finalize the inputted URD and to close the pop-up box.

participants completed the game to determine that no vocabulary deficits existed, which there were not.

In addition to the primary dataset, we collected data from 2 dementia caregivers as well. This data uses the same word scramble game as the original cohort and consisted of the same ordering and number of scrambled words. The difference was that the game was broken down into subsets of four words per day over 1 week period for a total of 2 weeks. Thus, both participants completed two full word scramble games over their participation time. The CAST app was pre-installed on an Android tablet and given to the dementia caregivers for the duration of their participation. Three meetings with the caregivers were made. First, to deliver the tablet and explain the study. Then, the second visit was to gather

the first word scramble games data and to discuss with the caregiver about the app. Lastly, the third meeting was to retrieve the tablet from the caregiver. The data collection was formatted in this manner for two reasons. First, we wanted to lessen the time required per day the participant would spend on this rather than their patient with dementia. Second, a short four word game mimics the ideal permanent deployment state. We note that due to scheduling issues with the participants, there are 4 missing data points for participant 2 as 1 out of the 28 days of combined participant gameplay was not completed.

3.3 Dataset Description

The 28 words used in the dataset are curated from seven categories: general words, edibles, items, actions, animals, flowers, and colors (Table 3.1). Note that Table 3.1 only lists 27 words as the general word 'hazardous' is used twice during gameplay with two different scrambles. That is, the first version (v1) appears near the beginning of the game while second version (v2) appears near the end. The primary difference between v1 and v2 is that v1 has the "ous" suffix unscrambled, i.e. remaining intact at the end of the word, while v2 includes the "ous" suffix in the scramble. Following, from the word dataset, the longest word length is 9 letters while the shortest length is 5. Furthermore, the standard deviation of the word length is 1.42, and with a mean word length of 6.89. The standard deviation of 1.42 word length shows a reasonable difference in word sizes, not showing bias to short or long words.

Four words are chosen to show different possible distributions of URD from the participants. Table 3.2 displays the chosen word, its corresponding scramble (i.e. the string of letters which can be rearranged to the original word), and the number of URD for each of the three difficulty categories. The category thresholds are discussed in Section 3.4.1. The four words chosen in Table 3.2 are "knock," "daffodil," "twilight," and "pistachio." Beginning with "knock," 38 of the 48 URD were categorized as Easy. For both "daffodil"

Table 3.1: The set of words curated for the CAST application's word scramble game. The categories are as follows: general words, edibles, items, actions, animals, flowers, and colors.

General	**Edibles**	**Items**	**Actions**	**Animals**	**Flowers**	**Colors**
hazardous	water	prize	check	manatee	daffodil	khaki
liberty	mustard	nickel	knock		lavender	ebony
quakes	avocado	pickup	defuse		jasmine	orange
bright	raspberry	gargoyle	harvest			
twilight	pistachio					
midnight						
brilliant						

and "twilight," they presented a split URD distribution between the Easy and Hard difficulties. The word "daffodil" has 19 Easy and 23 Hard ratings with only 4 Medium ratings. Similarly, the word "twilight" had 16 Easy ratings and 28 Hard ratings with only 4 Medium ratings. Finally, "pistachio" had 41 of the 46 URD categorized as Hard. While the clear categorizations of "knock" and "pistachio" are desired, the split distributions of the words "daffodil" and "twilight" pose a distinct categorization problem, which is addressed in Section 3.4.2.

Table 3.2: To illustrate URD differences between words, four words of various distributions were chosen. First, an Easy and Hard split distribution is shown with "daffodil" and "twilight". Second, "knock" shows a clear Easy URD grouping. Finally, "pistachio" depicts a definite Hard URD grouping.

Word	**Scramble**	**Easy**	**Medium**	**Hard**
knock	nkokc	**38**	5	5
daffodil	iodlffda	**19**	4	**23**
twilight	lgtwihit	**16**	4	**28**
pistachio	isihctoap	4	1	**41**

3.4 Data Preprocessing

Preprocessing the dataset is done in two parts. First, Section 3.4.1 discusses the how the thresholds for the Easy, Medium, and Hard categories are created. Second, Section 3.4.2 describes how a global difficulty is created for each word in the dataset. Both of these data preprocessing tasks will be handled by implementing a Rasch Model, a psychometric procedure for creating measurements from categorical data [4], for each task. Specifically, the Rasch model is capable of measuring participant ability and word difficulties, in conjunction and independently, on a log-odds scale. That is, a single word can be assigned a difficulty with respect to the URD from the entire participant cohort. Finally, the Rasch model is implemented using the Bigsteps Rasch model software.

3.4.1 Threshold Model

As mentioned in Section 3.3, a categorization problem exists as the URD do not agree for a subset of words. Thus, making it difficult to choose which of the three categories a word should belong in. Therefore, the first Rasch model, called the Threshold Model (TM), is created using the collective URD of the 28 words to address this problem. The TM measures the category probabilities with respect to the URD of the 28 words. Specifically, the TM will provide a two dimensional plot depicting the probability of response of a specific difficulty versus the person minus the item measure, i.e. the proficiency of the participants. The participants proficiency in the CAST word scramble game is calculated by the TM using a logistic function which produces values on a log-odds scale. This plot, called the category probabilities plot, allows us to visualize the the probabilities of any given rating from 1-10 and to justify which ratings belong to the Easy, Medium, and Hard categories. The discussed category probabilities plot is presented in Section 4.2.

3.4.2 Performance Model

The second Rasch model implementation is done using the data depicting whether the participant successfully unscrambled the word or not. This Performance Model (PM) is utilized for deciding the global difficulties to be assigned to each word in the dataset. That is, the PM maps the correct or incorrect responses (labeled 0 for incorrect and 1 for correct) to a singular difficulty value, which corresponds to a similar player ability scale, using a logistic function which outputs a log-odds value. To convert to the standard 1-10 scale of the URD from the log-odds scale, we map a 1 to the smallest log-odds value and a 10 to the largest log-odds value. Following, we calculate the equation of a line to do the conversion from the log-odds scale to the URD scale. Equation 3.1 is thus formulated and contains two variables, **y** and **x**. The **y** variable is the converted URD output and the **x** variable is the corresponding inputted log-odds value of the same word. For example, for the word "check," the PM calculates the log-odds value of -2.51. When inputted into Equation 3.1, the resulting URD rating is 3.18. We then proceed to round the result to the nearest integer to produce the global Rasch Model Difficulty (RMD), the second ground truth for this thesis, of "check" to be 3.

$$y = 0.9595x + 5.5864 \tag{3.1}$$

3.5 Qualitative Application Usage

As mentioned in Section 3.1, it was stated that the CAST app was designed to have a simplistic layout and to be intuitive to use. In order to verify this claim, we created a short qualitative survey for a small group of participants to fill out after completing a shortened version of the CAST word scramble game. We found 15 participants of which there were 5

female and 10 male with ages ranging from 19-38. Of the 15 participants, 6 are pursuing a bachelor's degree and 9 are pursuing a master's degree or higher. The CAST word scramble game was given to each participant individually with no explicit instructions on how to operate the app. The only instructions were to complete the four words in the game, then complete the short survey once finished. The goal is to let the participants decide on how intuitive the app was via the survey. The analysis of the survey will be done by examining quotes from the survey and by utilizing basic statistics such as the maximum, minimum, mean, and mode. The 9 survey questions are given below.

1. On a scale of 1 to 5 where 1 is "very slow" and 5 is "very fast," how well did the application respond to your interactions in a timely manner?

2. On a scale of 1 to 5 where 1 is "complex" and 5 is "very intuitive," please rate the usability (i.e. was it frustrating, easy to use) of the application?

3. On a scale of 1 to 5 where 1 is "confusing" and 5 is "straightforward," how was the design (i.e. appealing, easy to follow) of the user interface?

4. Were there any moments in which you did not know how to proceed in the application? (Yes or No)

5. If you answered yes in question 4, please explain the problem you encountered.

6. Did you feel that you needed to learn additional information to use this application? (Yes or No)

7. If you answered yes in question 6, please explain what information would have helped you during the application usage.

8. On a scale of 1 to 5 where 1 is "strongly disagree" and 5 is "strongly agree," do you think that most people would quickly and easily learn how to use this application?

9. Do you have any additional comments or suggestions?

3.6 Fuzzy Inference System

To analyze the dataset collected by the CAST word scramble game, we utilize an FIS, which is a supervised machine learning technique. The first portion of the FIS, the fuzzy logic, was developed by Lotfi Zadeh in 1973 [40]. Fuzzy logic provides the capability of traditional classification while incorporating noise from the data. The second part of the FIS, the inference system, is a structure which utilizes linguistic rules formulated in an IF-THEN manner, discussed in Section 3.6.3; as well as membership functions, described in Section 3.6.1. These membership functions are used to determine the degree for which an output belongs to a label. Combining the fuzzy logic and the inference system creates the FIS tool. The process of ingesting data starts with a data point being "fuzzified" and inputted into the hierarchical FIS, explained in Section 3.6.2, and the result is "defuzzified" into a readable value, in this case an Individualized Word Difficulty (IWD). This IWD output is compared to both the URD and RMD ground truths. The FIS is chosen as the linguistic variables and membership functions allow for semantic explainability [21, 29] to experts not in a technical domain. Finally, to construct the FIS, we used MatLab R2018a with the Fuzzy Logic Designer toolbox.

3.6.1 Membership Functions

Membership functions, which are created from an extension of the dataset features, are one of the two primary parts of the FIS. They utilize the features of the dataset and are divided into three portions, the label, distribution form, and distribution parameters. Specifically, the label represents how the feature is treated when used with the corresponding linguistic rule. The distribution forms, while using MatLab's Fuzzy Logic Designer toolbox, can be chosen from a list of eleven distributions such as Gaussian, Triangular, or Sigmoid. Finally, Each distribution form has corresponding parameters which are specific to what form is chosen. For example, a Gaussian distribution requires a mean and standard deviation pa-

rameter.

The features extracted from the CAST word scramble game which are used for the membership functions are shown in Table 3.3. The five features are Time Taken, Number of Guesses, Number of Letters, Was Skipped, and Degree of Scramble. These features were selected to balance low complexity while still maintaining meaningful feature information. Furthermore, the extracted features take into account both user and gameplay aspects. That is, the features Time Taken and Number of Guesses correspond directly to an individual user. Conversely, the features Number of Letters and Degree of Scramble relate to the word during gameplay.

Table 3.3: The features extracted from gameplay of the CAST word scramble game. Five features were chosen for low complexity while still extracting meaningful information. The features are Time Taken, Number of Guesses, Number of Letters, Was Skipped, and Degree of Scramble.

Feature	**Description**	**Type**
Time Taken	Time taken in seconds to correctly finish or skip the task.	Integer
Number of Guesses	Number of guesses for a single task.	Integer
Number of Letters	Number of letters in the scrambled word.	Integer
Was Skipped	Whether the task was skipped or not.	Binary
Degree of Scramble	Degree of scramble of the word.	Float

Of the five features, only the Degree of Scramble feature is not immediately intuitive. Degree of scramble, a statistic created specifically for this thesis, is calculated by comparing a word W and its corresponding permutation (scramble) P. The comparison is done on a letter by letter basis to create a multitude of tuples. The Degree of Scramble is then the aggregate of each tuple of W and P for which the tuple does not share the same letter at the index i. Equation 3.2 represents this process as a quickly converging series where $I(w,p)$ is shown with Equation 3.3. So, every ith letter of a word W and permutation P is inputted from Equation 3.2 which then utilizes Equation 3.3 and it returns a 1 if the letters are different or a 0, otherwise. This equation was created as we hypothesize that letters at

the beginning of the word, such as the prefix, provide more information for unscrambling a word rather than perhaps letters in the middle of a word. A Degree of Scramble value approaching 1 shows a complete scramble where there are no matching letters between W and P, while no scramble produces a 0. An immediate limitation to this metric is that a letter shift left or right creates a complete scramble while still being relatively simple to unscramble. This is accounted for with and no such scramble existing in the dataset.

As this is a new statistic, we compare it against a well documented measure called the Hamming distance [16]. The Hamming distance is a metric which measures the number of substitutions it takes to go from one string of letters to another. We calculated the Hamming distance for all 28 words using their corresponding scrambles and divide the results by the length of the word. As an example, the word "prize" scrambled as "repiz" has a Hamming distance of 5, divided by 5 gives the value 1.0, while our metric gives the value 0.97. Following, we compare our metric with the Hamming distance by calculating the Pearson correlation. We report statistically significant correlation between our Degree of Scramble metric and the Hamming distance ($r=0.47$, $p<0.05$).

$$S(W,P) = \sum_{i=1}^{n} 1/2^i * I(w,p) \tag{3.2}$$

$$I(w,p) = \begin{cases} w \neq p & 1 \\ \text{otherwise} & 0 \end{cases} \tag{3.3}$$

While heuristically designing the membership functions shown in Tables 3.4 and 3.5, we make the assumption that the data generally follow a Gaussian distribution. Therefore, a majority of the membership functions use a Gaussian form and all the corresponding parameters are evenly spaced on the input plane. Two separate tables exist for the member-

ship functions, the first representing the input features from the dataset (Table 3.4), and the second representing the output of the two intermediary FIS and the final IWD FIS (Table 3.5). The intermediary FIS in the hierarchical construction is further discussed in Section 3.6.2.

Table 3.4: The dataset features represented as membership functions. Each membership function consists of three parts. The linguistic label to be represented in a rule, a distribution form, and the distribution parameter.

Input Features	Membership Functions		
	Label	Form	Parameter
Number of Guesses	Low	Gaussian	[1.699 0]
	Medium	Gaussian	[1.699 5]
	High	Gaussian	[1.699 10]
Time Taken	Short	Gaussian	[10.19 0]
	Medium	Gaussian	[10.19 30]
	Long	Gaussian	[10.19 60]
Was Skipped	True	Triangular	[-.01 0 .01]
	False	Triangular	[.99 1 1.01]
Length of Word	Short	Gaussian	[.85 5]
	Long	Sigmoid	[2.38 6.53]
	Very Long	Gaussian	[.85 10]
Degree of Scramble	Low	Gaussian	[.1699 0]
	High	Sigmoid	[.1699 .5]
	Very High	Gaussian	[.1699 1]

3.6.2 Hierarchical Construction

As discussed in Section 3.6.1, the five extracted features relate to both individual users and to gameplay as well. Therefore, a two stage hierarchical construction of the FIS was utilized that allows for differentiation between the user and the gameplay before making a final IWD classification. Shown in Figure 3.3, the left column accommodates the five input features, the center column contains two intermediary FIS, and the right column consists of the single IWD output FIS. The final output of the system will be a difficulty rating of either Easy, Medium, or Hard on a per word basis.

Table 3.5: The FIS output membership functions. Each membership function consists of three parts. The linguistic label to be represented in a rule, a distribution form, and the distribution parameter.

FIS Outputs	Membership Functions		
	Label	Form	Parameters
User Effort	Low	Gaussian	[.1699 0]
	Medium	Gaussian	[.1699 .5]
	High	Gaussian	[.1699 1]
Complexity of Word	Low	Gaussian	[2.123 0]
	Medium	Gaussian	[2.123 .5]
	High	Gaussian	[2.123 1]
IWD	Easy	Gaussian	[1 1.6]
	Medium	Gaussian	[1 4.6]
	Hard	Gaussian	[1.5 8.9]

The two intermediary FIS in stage one are designed to produce information from the user and the gameplay, respectively. That is, the first FIS in stage one, the User Effort (UE) FIS, ingests the Time Taken and Number of Guesses features. The output of this FIS is a float value from 0 to 1 that represents how much effort a user gave for a given word. Following, the second FIS in stage one, the Complexity of Word (CoW) FIS, takes in the Number of Letters and Degree of Scramble features. Thus, outputting a float value between 0 and 1 which represents the aspects of the word scrambling task. The output of both stage one FIS, in conjunction with Was Skipped feature, are directed to the single stage two IWD FIS. The reason the Was Skipped feature is not used in either of the stage one FIS is because this feature is a direct indicator of task difficulty. Finally, the membership function labels, forms, and parameters for all three FIS in the system can be seen in Table 3.5.

3.6.3 Rule System

The two stage hierarchical FIS system consists of 16 heuristically created linguistic rules. The rules were created by examining early data and discussing with clinical collaborators. Each rule is structured in an IF-THEN manner. That is, the **IF** clause is the antecedent and

the **THEN** clause is the consequent. The rules are listed in Tables 3.6, 3.7, and 3.8 for the UE, CoW, and IWD FIS, respectively. Examining the UE rules from Table 3.6, the first rule can be deciphered as *IF* the Number of Guesses is **Low** *AND* the Time Taken is **Short** *THEN* the User Effort is **Low**. This rule is intuitive in that if a user only guesses 3 times and spends 20 seconds on the task, it follows that a low amount of effort was used.

All 16 rules follow the same structure of utilizing the membership functions and logical construction. We illustrate this process with Figure 3.4. The figure shows the stage two IWD FIS with the inputs from the Was Skipped feature and the stage one CoW and UE FIS. In this manually curated example, we use the word "water," with the scrambled version of "tarew." The values of 5 and 0.66 were used for the features Length of Word and Degree of Scramble, respectively, in the CoW FIS. This resulted in the CoW FIS outputting the value of 0.17. Following, we assume the participant attempted 2 guesses with 15 seconds of time spent on the task. These two inputs are inserted into the UE FIS which outputs the value 0.35. Finally, the results from both stage one FIS and a 0 for the Was Skipped feature, i.e. the participant successfully unscrambled the word, were inputted into the stage two IWD FIS. From Figure 3.4, we can see six rows where each correspond to the rule listed in Table 3.8. After the rules utilize the membership functions outputs, the far right column shows the resulting curves for each rule which will be used to calculate the final IWD. Aggregating the six curves and taking the centroid, as shown in the bottom right of the figure with the red line representing the centroic, gives the final IWD of 3.64.

3.7 Genetic Algorithms

There are a total of 23 membership functions in the two stage hierarchical FIS, 21 of which have floating point parameter values allowing for an extremely large number of combina-

Table 3.6: The stage one UE FIS. The antecedent, which utilizes the Number of Guesses and Time Taken features, is seen in the first two columns. The consequent, which outputs the User Effort value, is seen in the third column. The Number of Guesses labels are Low, Medium, and High. Following, the Time Taken labels are Short, Medium, and Long. Finally, the User Effort labels are Low, Medium, and High.

Antecedent						**Consequent**		
Number of Guesses			**Time Taken**			**User Effort**		
L	M	H	S	M	L	L	M	H
X			X			X		
	X	X		X			X	
		X			X			X
			X			X		
					X			X

Table 3.7: The stage one CoW FIS. The antecedent, which utilizes the Length of Word and Degree of Scramble features, is seen in the first two columns. The consequent, which outputs the Complexity of Word value, is seen in the third column. The Length of Word labels are Short, Long, and Very Long. Following, the Degree of Scramble labels are Low, Medium, and High. Finally, the Complexity of Word labels are Low, Medium, and High.

Antecedent						**Consequent**		
Length of Word			**Degree of Scramble**			**Complexity of Word**		
S	L	VL	L	M	H	L	M	H
X			X				X	
	X		X					X
					X	X		
		X		X				X
X				X	X	X		

tions. Thus, it is not feasible to manually adjust the membership function parameters to find a near optimal solution. To address this problem, we implemented GA as a way of automatically finding a globally optimized solution. Specifically, 21 of the 23 membership functions shown in Tables 3.4 and 3.5 are updated with the GA. The two membership functions excluded from the GA are the Was Skipped membership functions. As these are boolean values, it does not make sense to modify their parameters.

The GA operates in a cyclic fashion that terminates once an optimization condition is met. Figure 3.5 depicts this cycle which begins with an initialization step. Next, a population of chromosomes, i.e. a set containing the membership function values, are created

Table 3.8: The stage two IWD FIS. The antecedent, which utilizes the User Effort, Complexity of Word, and Was Skipped outputs, is seen in the first three columns. The consequent, which outputs the Individualized Word Difficulty value, is seen in the third column. The User Effort labels are Low, Medium, and High. Following, the Complexity of Word labels are Low, Medium, and High. Next, The Was Skipped labels are True and False, Finally, the Individualized Word Difficulty labels are Easy, Medium, and Hard.

Antecedent								**Consequent**		
User Effort			**Complexity of Word**			**Was Skipped**		**IWD**		
L	M	H	L	M	H	T	F	E	M	H
X							X	X		
X	X		X	X			X	X	X	
X					X	X				X
		X								X
X							X	X		
X					X		X		X	

while taking into account the problems constraints. Parent selection is then performed, which in this case is a stochastic uniform selection. That is, the parents are sorted based on their fitness function result, where the fitness function is the sum of the squared error function, and a stochastically uniform step is taken to choose the subset of parents. With the pool of parents created, crossover occurs, which combines two parent chromosomes to create a new chromosome. Similarly, these new chromosomes also go through a mutation step which randomly modifies a positions value in the chromosome while remaining within the problems constraints. Once both crossover and mutation are completed, a new child pool is created. Following, survivor selection, similar to parent selection, is then performed. The cycle then returns to the fitness check stage that will check the cycle termination conditions. Finally, the GA were implemented using MatLab's Global Optimization toolbox. The settings used to initialize the GA are shown below.

- **Population Size** = *200* : Initial population size of a chromosome.
- **Creation Function** = *Uniform* : Randomly generates the initial population by sampling from a uniform distribution.
- **Scaling Function** = *Rank* : Scales the scores of individual chromosomes in a sorted

order where the first chromosome is the most fit with respect to the sum of squares error function.

- **Selection Function** = *Stochastic Uniform* : The selection of the parents and children is done by sorting the chromosomes by fitness. Then, a subset of chromosomes are selected by randomly stepping through the list with uniform probabilities.

- **Mutation Function** = *Adaptive Feasible* : Checks for feasible chromosomes are chosen for reproduction by randomly updating a position in a chromosome based upon the last generation.

- **Crossover Function** = *Scattered* : Randomly creates a vector of the size of the chromosome containing 1's and 0's. That is, the 1's takes the value from the position in the first parent chromosome and puts it in the child's while the 0's performs the same action with the second parent.

- **Upper and Lower Bounds** = *Tables 3.4 and 3.5 Parameter Column* : The upper and lower bounds of the membership functions for the FIS are used to prevent unexpected results from occurring.

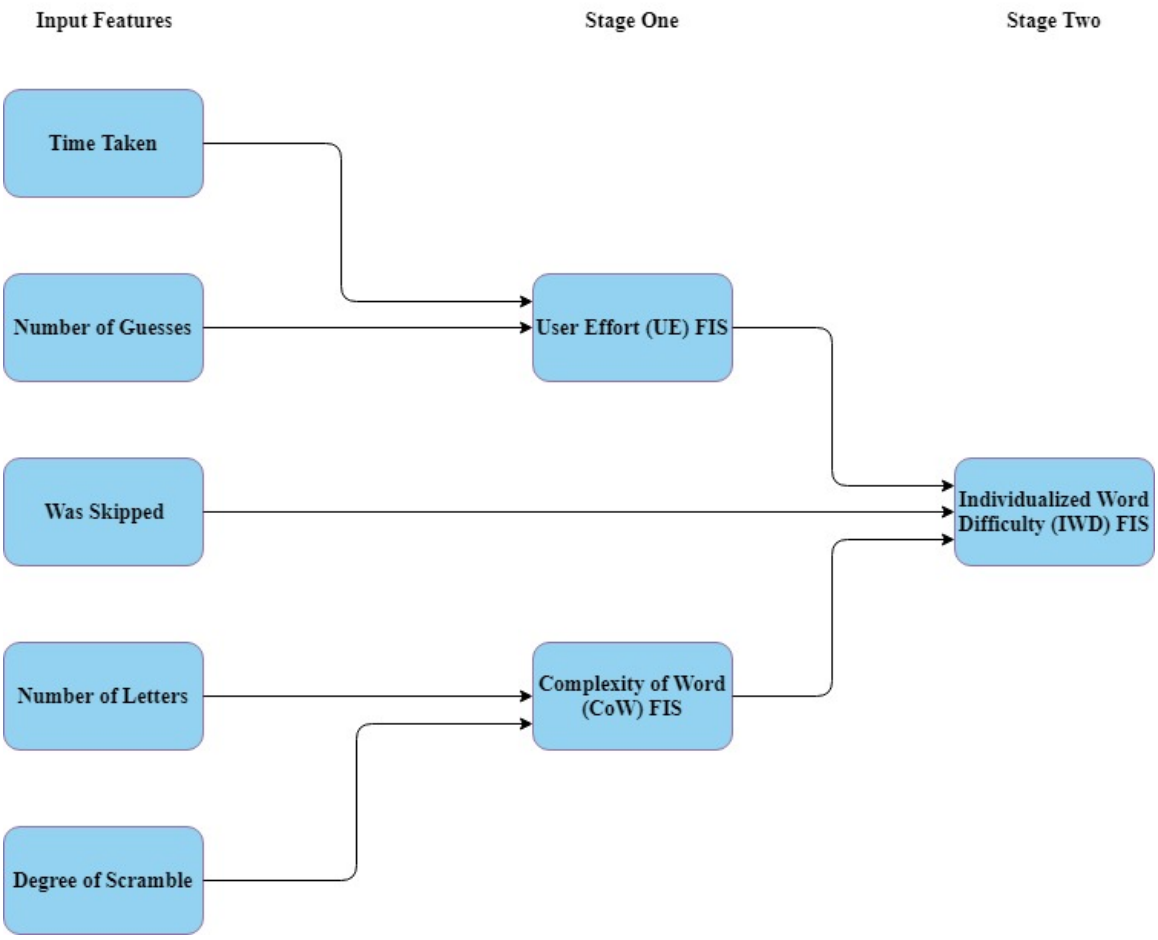

Figure 3.3: The hierarchical flow of the two stage FIS system. Stage one first takes in as inputs the Time Taken and Number of Guesses features for the UE FIS. The features Length of Word and Degree of Scramble are used as inputs to the CoW FIS. Finally, stage two ingests the output of the UE and CoW FIS as well as the Was Skipped feature to output the final IWD of a word.

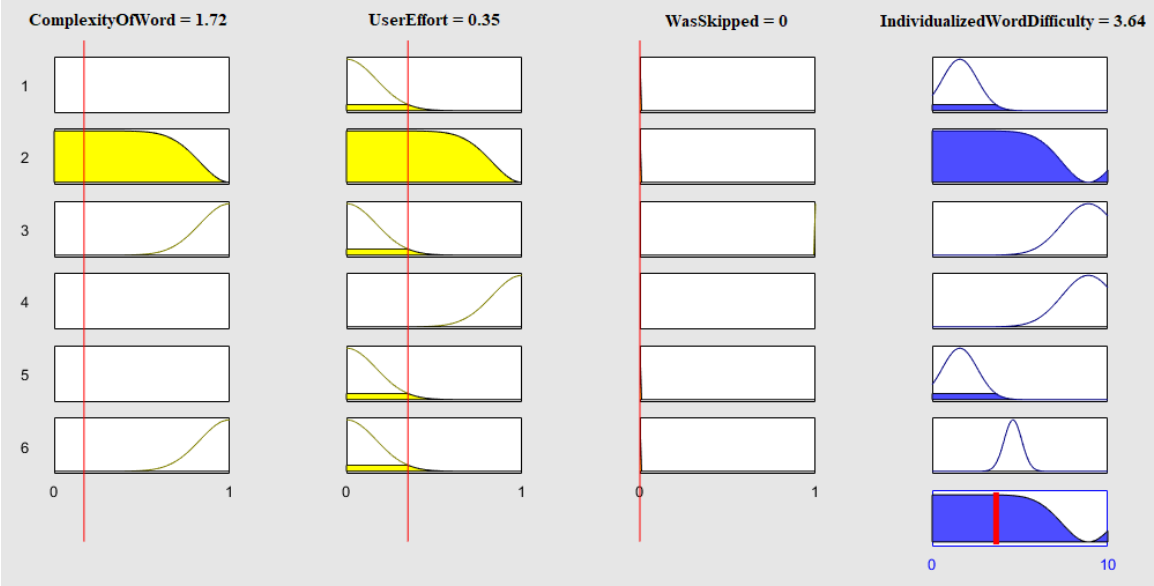

Figure 3.4: The graphical visualization of the IWD FIS during execution. There are six rows which correspond to the six IWD rules. Following, the first three columns show the CoW and UE FIS inputs as well as the Was Skipped feature input, respectively, and how they visually interact with each rule. The fourth column is the consequent of each rule where the aggregate of the six consequents is shown at the bottom of the fourth column in which the centroid is taken for the final IWD output.

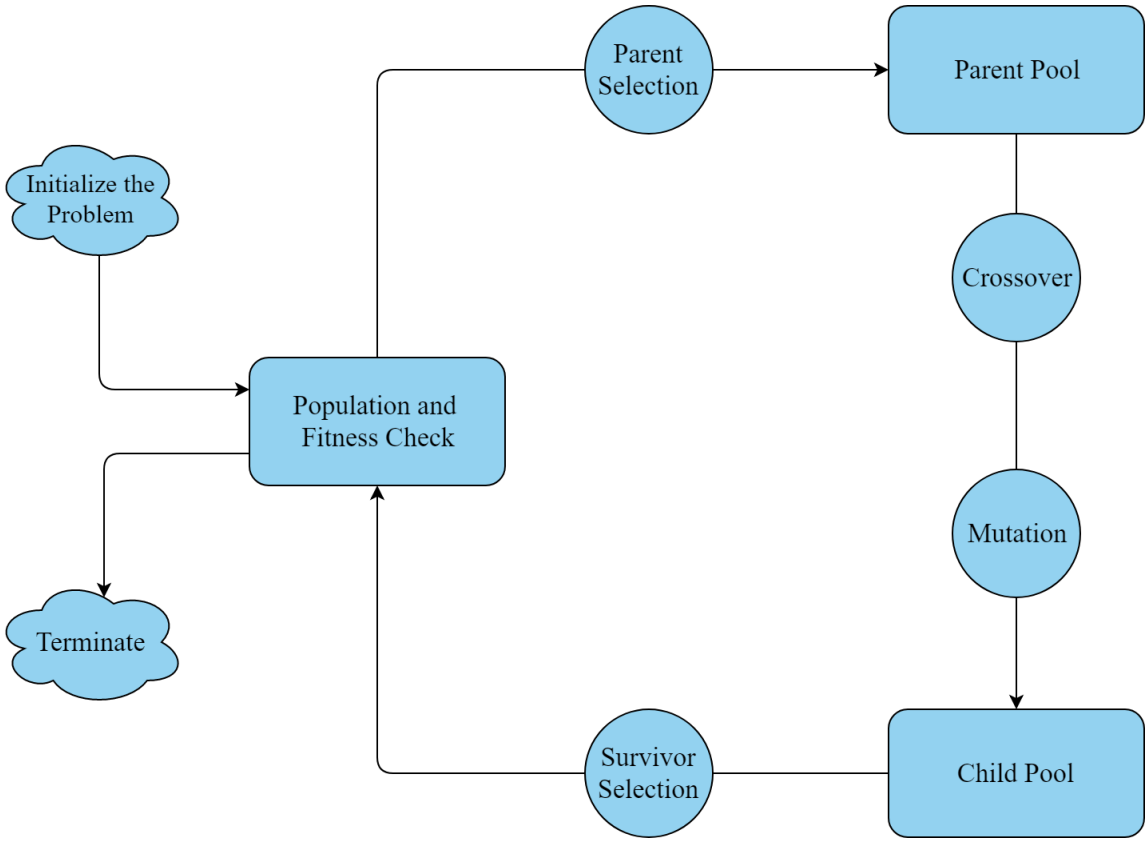

Figure 3.5: The cycle which the GA performs. Initialization begins where a population is created within the problem constraints. Next, parent selection is performed to keep a subset of chromosomes as determined by the sum of the squared error function. With the parent pool created, crossover and mutation, both methods of creating new unseen child chromosomes, is performed. Survivor selection is performed on the child pool to have a final population, which is checked via the fitness function for the termination conditions.

4

Results

This chapter discusses the results of the qualitative CAST usability survey, the Rasch model data preprocessing, the heuristically created FIS, and the GA optimized FIS. Specifically, Section 4.1 presents the results of the qualitative survey using basic statistics such as the maximum, minimum, and mean. Then, Section 4.2 describes how the Rasch TM created the thresholds used for the Easy, Medium, and Hard categories. Section 4.2 also discusses the Rasch PM and how it created the global difficulties for each word in the dataset. Following, Section 4.3 shows the results of the heuristically created FIS using the precision, recall, and F1 score performance metrics. Similarly, Section 4.4 presents the results of the GA improved FIS using the same performance metrics as the heuristically made FIS. Finally, Section 4.5 shows the results of the longitudinal study using the CAST word scramble game with the 2 dementia caregiver participants.

4.1 Qualitative CAST Usability Survey

The results of the qualitative usability survey comes in the form of basic statistics which include the maximum, minimum, mean, and mode. There are four questions from the survey that we will be performing the statistics on, which are 1 (responsiveness), 2 (intuitiveness),

3 (ease of use), and 8 (learnability). Furthermore, we present the number "Yes" responses from questions 4 (did not know how to proceed) and 6 (additional information needed) regarding confusion or lack of information (please see Section 3.5 for the survey questions). Shown in Table 4.1, the results of the survey are promising. The minimum for any of the questions is a 3 out of 5 where the mode for all four questions is a 5 out of 5. Furthermore, for both questions 2 (intuitiveness) and 3 (ease of use) which have a 3 for a minimum, only two participants gave this response (one for each question). Following, there were only 3 participants that answered "Yes" for question 4 (did not know how to proceed) and 0 participants answered "Yes" for question 6 (additional information needed).

Table 4.1: The results of questions 1, 2, 3, and 8 from the qualitative CAST usability survey. The maximum, minimum, mean, and mode are presented. The questions can be referenced in Section 3.5. These four questions represent the responsiveness, design, and intuitiveness of the CAST word scramble game.

	CAST Usability Survey Questions			
	Question 1	Question 2	Question 3	Question 8
Maximum	5	5	5	5
Minimum	4	3	3	4
Mean	4.73	**4.47**	**4.47**	**4.87**
Mode	5	**5**	**5**	**5**

4.2 Rasch Model Analysis

As discussed in Section 3.4.1, the thresholds are determined by using the Rasch TM. The TM produces the Category Probabilities plot shown in Figure 4.1. The plot shows the ten possible ratings from 0-9 which are shifted by 1 for presentation purposes, but actually reflect the CAST 1-10 ratings. The probability of response for any given rating are plotted against the person minus the item measure, i.e. the proficiency of the participant. Finally, a "*" represents multiple possible ratings in any position on the figure.

Visual inspection of the graph shows two clear large categories and one smaller category between them. From the first large category, we can see a high probability of 0.6 through 0.8 for the 0 rating. We also see the ratings 1, 2, and 3 with lower probabilities between approximately 0.1 through 0.3, but still in the same participant proficiency range which the 0 rating exists. Following, from the second large category, we see another high probability of 0.5 through 0.9 for the 9 rating. Also, the ratings of 6, 7, and 8, as well as the mixed rating of 5, follow the same general lower probabilities as the first large category of approximately 0.1 through 0.2. Finally, the last category is shown where the probabilities for the 0 and 9 ratings enter the same approximate range as the remaining ratings and therefore, the rating of 4 is chosen. Using these thresholds from the Rasch TM, the rating categories of 0-3, 4, and 5-9, converted to 1-4, 5, and 6-10 for the CAST app are chosen for the Easy, Medium, and Hard categories, respectively.

To address the problem of determining word difficulty placement, as discussed in Section 3.4.2, the Rasch PM was implemented. The PM model works to remove the URD split rating bias from a subset of words in the dataset. This, in turn, allows for assigning a global difficulty to each word, creating the RMD ground truth. To achieve this, the PM produced a log-odds difficulty rating scale for each word in the dataset, from which Equation 3.1 was created to convert the values to the CAST rating scale of 1-10. After converting the difficulties to the CAST rating scale and thus, creating the RMD ground truth, we obtain a finer distribution of word difficulties as shown in Figure 4.2. The presented histogram shows that while there is slight negative skew in the overall RMD ratings favoring the Hard difficulty, words do exist for both the Easy and Medium difficulties as well. Table 4.2 shows the RMD ratings for each word in the dataset.

Table 4.2: The rounded RMD ratings produced from the Rasch PM for each word in the dataset. We can see that a majority of words are within the Hard (6-10) category. Though, both the Easy (1-4) and Medium (5) categories both contain words meaning that the category is sufficient to use.

Words per each Rounded RMD Rating Bin					
3	4	5	6	7	8
check	**hazardous v1**	water	ebony	**manatee**	**pistachio**
knock	pickup	prize	nickel	khaki	
midnight		daffodil	orange	liberty	
		bright	quakes	mustard	
		avocado	defuse	brilliant	
		twilight	jasmine	lavender	
		raspberry	harvest	**gargoyle**	
			hazardous v2		

4.3 Individualized Word Difficulty Predictions Using FIS

Using the calculated thresholds to determine the Easy, Medium, and Hard difficulties category size, the IWD from the heuristically created FIS is tested against the URD and RMD ground truths. Table 4.3 utilizes the resubstitution method for both ground truths while Table 4.4 does the comparison using the leave-one-out method. The performance metrics of Precision, Recall, and F1 score are implemented for analysis of the FIS IWD results. We note that in both Tables 4.3 and 4.4, the URD are word dependent. That is, unlike how each word with the global RMD rating will not change in this thesis, the URD can have multiple ratings for a single word. For example, the word "water" could have both a rating of 2 and 8, placing it in both the Easy and Hard category, respectively.

Examining Table 4.3, we can see that the Medium category has the values of 0.20, 0.08, and 0.11 for Precision, Recall, and F1 score, respectively, for the RMD ground truth. Similarly, the URD ground truths Medium category has the values of 0.13, 0.18, and 0.15 for Precision, Recall, and F1 score, respectively. This clearly shows an innate difficulty in classifying the Medium category. Conversely, both the Easy and Hard categories greatly outperform the Medium category with the resubstitution method. The RMD ground truth

has Precision, Recall, and F1 score of 0.30, 0.86, and 0.45 for Easy and 0.77, 0.53, and 0.63 for Hard. Similarly, the URD ground truth has Precision, Recall, and F1 score of 0.68, 0.95, and 0.79 for Easy and 0.94, 0.66, and 0.77 for Hard. There were 459 easy and 484 hard out of 1,320 total correct classifications for the URD resubstitution shown in Table 4.3. This supports are initial findings of the participant split ratings for a subset of words. We next examine Table 4.4 to see that the URD ground truth IWD comparison performs relatively well for the leave-one-out method. This is shown by the Precision, Recall, and F1 score of 0.94, 1.00, and 0.97 for Easy and 1.00, 1.00, and 1.00 for Hard.

Table 4.3: Performance of the heuristically created FIS using the Precision, Recall, and F1 score metrics. This table uses the resubstitution method to compare the RMD and URD ground truths to the IWD output.

	Resubstitution					
	RMD			URD		
	Easy	Medium	Hard	Easy	Medium	Hard
Precision	0.30	0.20	0.77	0.68	0.13	**0.94**
Recall	0.86	0.08	0.53	**0.95**	0.18	0.66
F1 score	**0.45**	**0.11**	**0.63**	0.79	0.15	0.77

Table 4.4: Performance of the heuristically created FIS using the Precision, Recall, and F1 score metrics. This table uses the leave-one-out method to compare the RMD and URD ground truths to the IWD output.

	Leave-One-Out					
	RMD			URD		
	Easy	Medium	Hard	Easy	Medium	Hard
Precision	0.31	0.00	0.82	**0.94**	1.00	**1.00**
Recall	1.00	0.00	0.56	**1.00**	0.50	**1.00**
F1 score	0.47	0.00	0.67	**0.97**	0.67	**1.00**

4.4 Individualized Word Difficulty Predictions Using GA

With the GA implemented to improve the membership functions of the heuristically created FIS, we conducted the same performance tests as were done in the previous section.

We present Tables 4.5 and 4.6 which show the resubstitution and leave-one-out methods, respectively. Both tables follow the same procedure in which Precision, Recall, and F1 score are used to analyze the performance of the GA improved FIS IWD output against the RMD and URD ground truths. We note that multiple GA models were trained with different settings such as no parameter bounds, different ground truth training comparisons (i.e. rounded and unrounded RMD), and a combination of both. Though, the presented results come from the best performing model.

Examining Table 4.5, we can see that the GA improved FIS had a positive effect on the Medium category with resubstitution. Using the RMD ground truth, the Medium category had values of 0.48, 0.45, and 0.47 for Precision, Recall, and F1 score, respectively. Though still low, it is a vast improvement over the heuristically created FIS. Following, the resubstitution Hard category of the URD ground truth improved with Precision, Recall, and F1 score of 0.86, 0.85, and 0.86. We also see the resubstitution URD Medium category still performing poorly (precision 0.15, recall 0.53, F1 score 0.23) due to the user difficulty rating bias. Next, Table 4.6 shows the RMD leave-one-out method's Hard category (precision 0.71, recall 0.63, F1 score 0.67) had poorer performance than the URD (precision 0.79, recall 1.00, F1 score 0.88). Overall, the GA optimization of the FIS membership functions greatly improved the performance of the FIS.

Table 4.5: Performance of the GA improved FIS using the Precision, Recall, and F1 score metrics. This table uses the resubstitution method to compare the RMD and URD ground truths to the IWD output.

	Resubstitution					
	RMD			URD		
	Easy	Medium	Hard	Easy	Medium	Hard
Precision	0.47	**0.48**	0.75	0.85	**0.15**	**0.86**
Recall	0.54	**0.45**	0.73	0.49	**0.53**	**0.85**
F1 score	0.50	**0.47**	0.74	0.62	**0.23**	**0.86**

Table 4.6: Performance of the GA improved FIS using the Precision, Recall, and F1 score metrics. This table uses the leave-one-out method to compare the RMD and URD ground truths to the IWD output.

	Leave-One-Out					
	RMD			URD		
	Easy	Medium	Hard	Easy	Medium	Hard
Precision	1.00	0.40	**0.71**	1.00	0.20	**0.79**
Recall	0.80	0.57	**0.63**	0.27	1.00	**1.00**
F1 score	0.89	0.47	**0.67**	0.45	0.33	**0.88**

4.5 Longitudinal Caregiver Deployment Results

As the original participant cohort was used for testing the creation of the FIS and for the improvements with the GA, we will only use the GA improved FIS to examine the results of the longitudinal study of the 2 dementia caregivers. Examining the four Figures 4.3, 4.4, 4.5, and 4.6, we can see the temporal results of the CAST word scramble game, i.e. the results of the 2 participants' games from week 1 and 2, respectively. The figures show the IWD and URD for weeks 1 and 2 for a subset of words in the dataset. This subset was chosen to illustrate the interesting phenomenon that occurs when the CAST word scramble game is deployed over time. That is, the changes and adaptations we see in the participants over multiple uses. We note that a majority of the words decline in IWD from week 1 to week 2 of 21/28 and 14/24 for participants 1 and 2, respectively. Furthermore the URD also declined with 15/28 and 8/24 URD being lower from week 1 to week 2 for participants 1 and 2, respectively, where the remainder of ratings were approximately the same. These results along with quotes from the two caregiver participants will be discussed in Section 5.5.

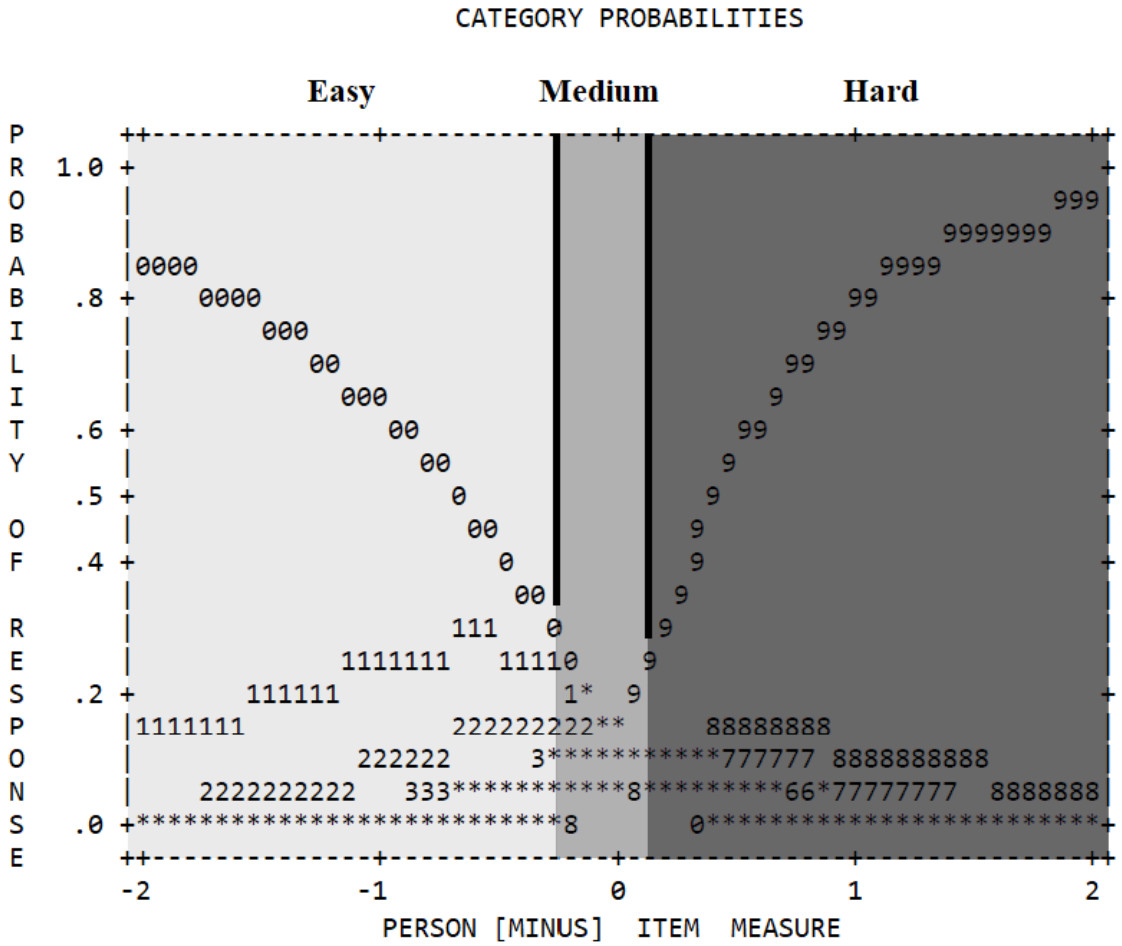

Figure 4.1: To determine the thresholds for the Easy, Medium, and Hard categories, we utilize the category probabilities produced from the Rasch TM. The plot shows the probability of response on the y-axis and the person minus the item score, i.e. the proficiency of the participant, on the x-axis. We determine the thresholds for each of the three categories by visually examining the probabilities of response for each possible threshold. These are shown as lines from 0-9 (converted from the word scramble game's 1-10 scale) and the "*" represents multiple possible ratings.

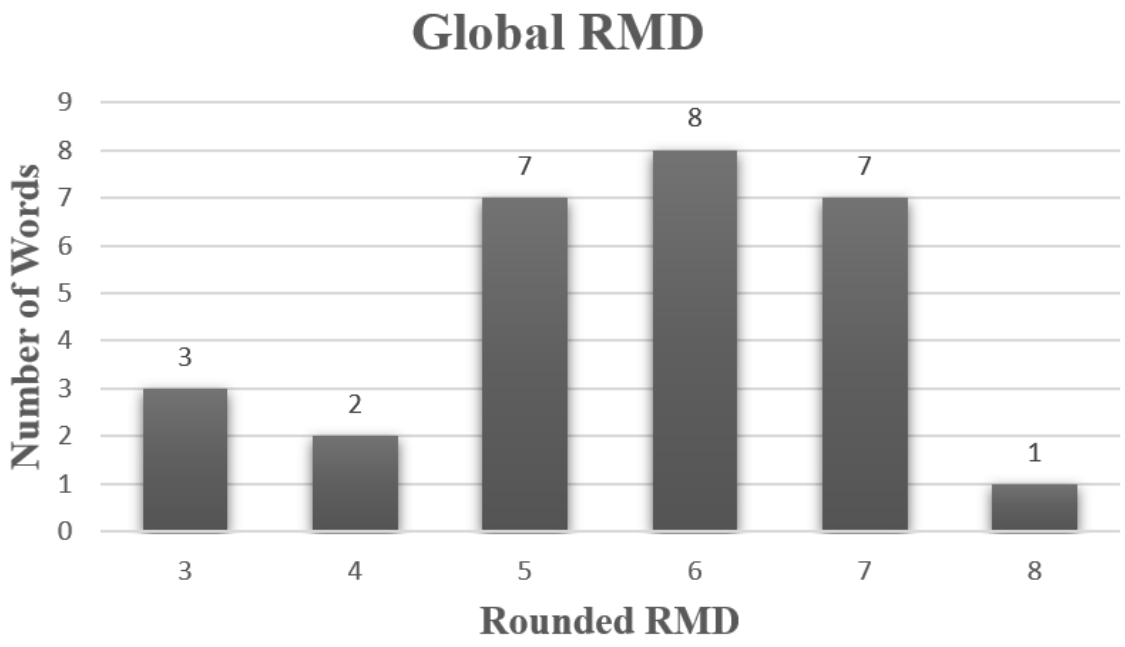

Figure 4.2: Histogram of the global RMD of each word produced from the Rasch PM. The y-axis shows the number of words per bin while the x-axis depicts the RMD bins, which can range from 1-10. From the figure, we can see a slight negative skew of the RMD ratings for the Hard (6-10) category. Though, there are words which exist for the Easy (1-4) and Medium (5) category as well.

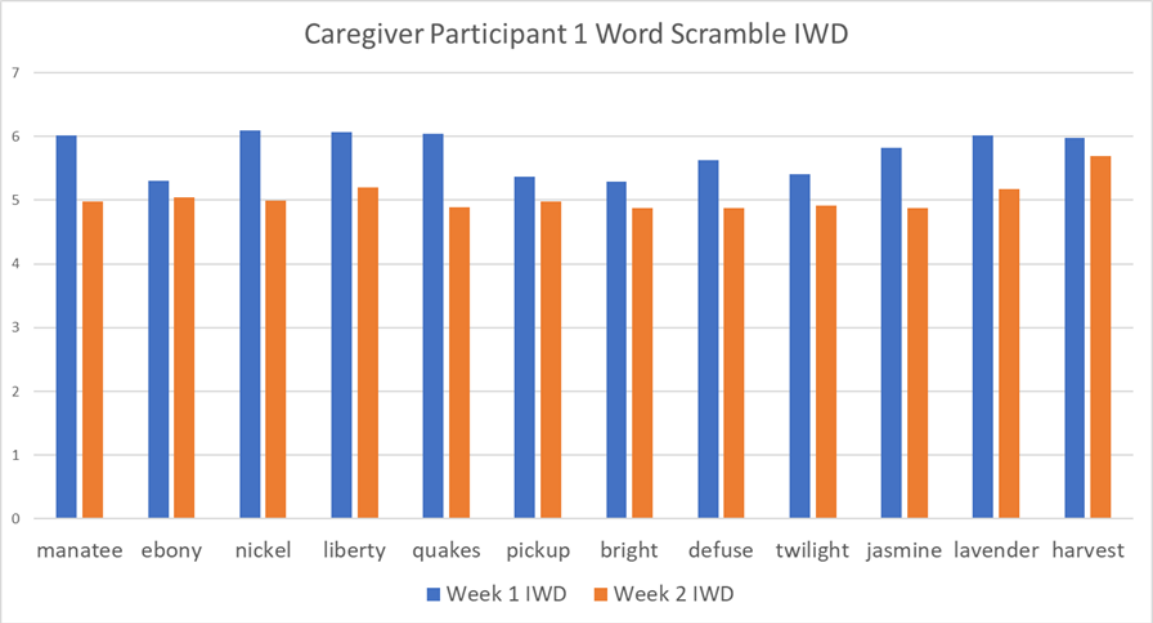

Figure 4.3: A subset of caregiver participant 1's words from the longitudinal study. The histogram shows the IWD of the words for both week 1 and 2. We can see that after only 2 iterations of gameplay that the IWD begins to lower.

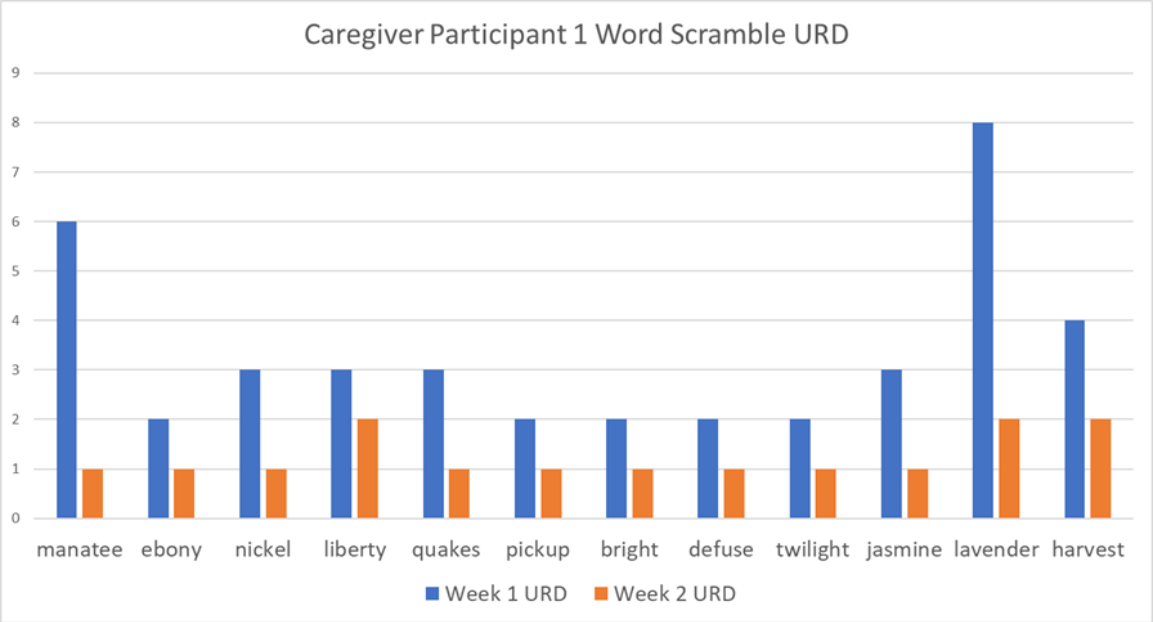

Figure 4.4: A subset of caregiver participant 1's words from the longitudinal study. The histogram shows the URD of the words for both week 1 and 2. We can see that after only 2 iterations of gameplay that the URD changes considerably.

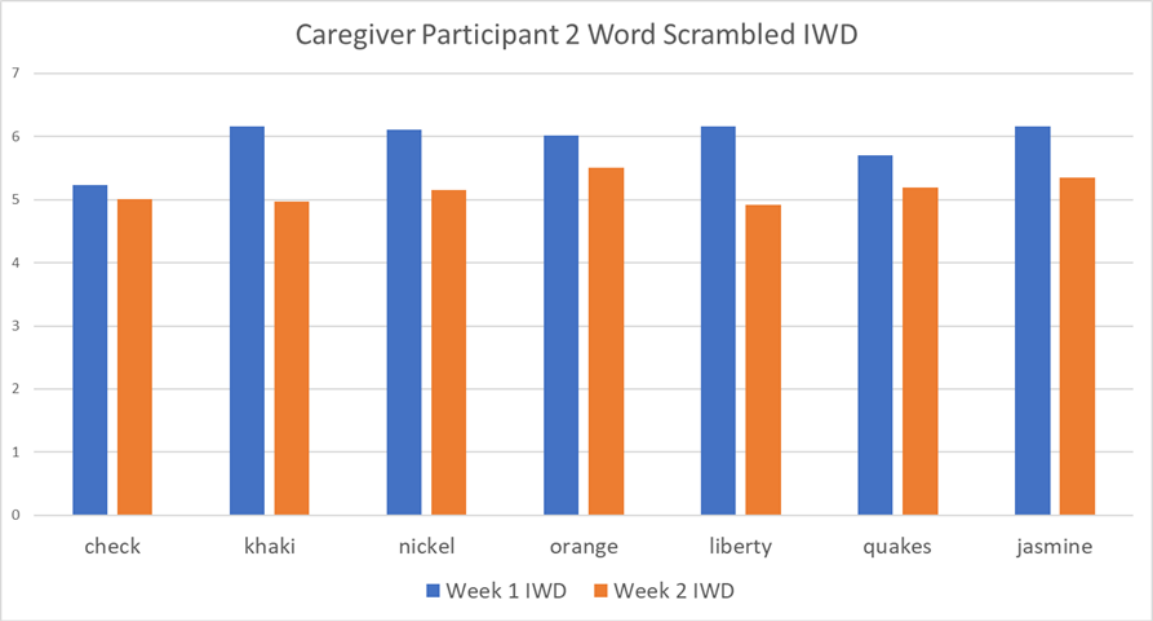

Figure 4.5: A subset of caregiver participant 2's words from the longitudinal study. The histogram shows the IWD of the words for both week 1 and 2. We can see that after only 2 iterations of gameplay that the IWD begins to lower.

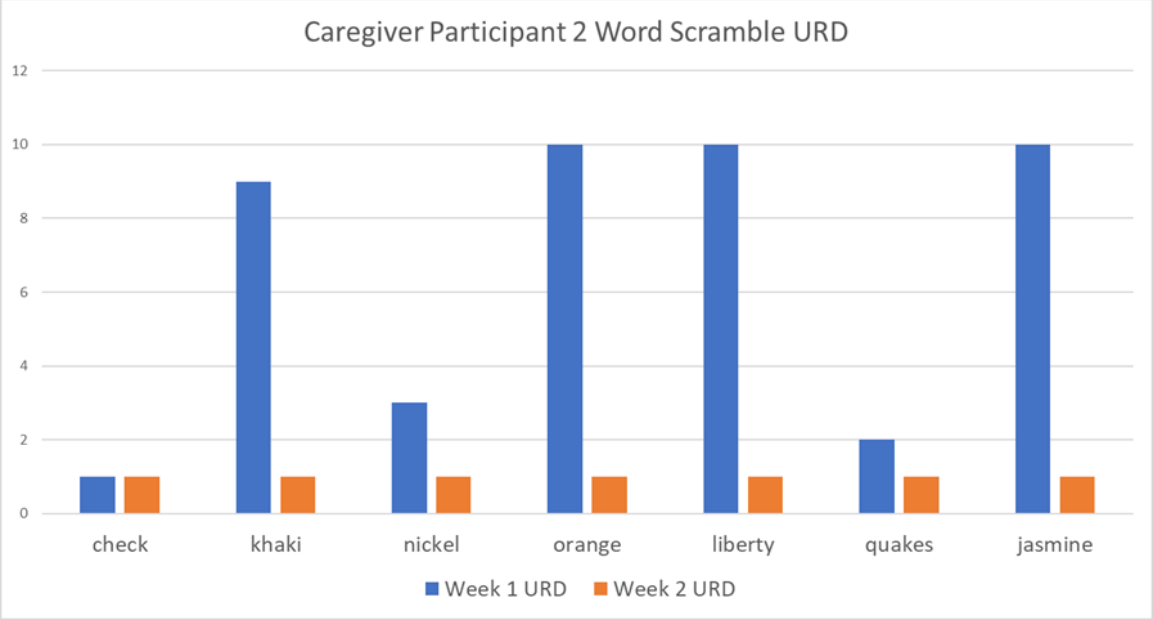

Figure 4.6: A subset of caregiver participant 2's words from the longitudinal study. The histogram shows the URD of the words for both week 1 and 2. We can see that after only 2 iterations of gameplay that the URD changes considerably.

5

Discussion

In this chapter, we analyze the word set chosen for the CAST app as well as the results from both the heuristically created FIS and the GA improved FIS. First, Section 5.1 analyzes the results of the qualitative CAST usability survey and presents multiple written responses from the opened ended questions as well. Following, Section 5.2 discusses the word set in detail with respect to both the URD and RMD ratings. Also, we discuss the usefulness of discriminatory words and how they relate to the goal of this study. Next, Section 5.3 explains the results of the heuristically created FIS. FThen, Section 5.4 discusses the results of the GA improved FIS. Furthermore, this section provides the decision on whether to use the URD or RMD ground truth label for future endeavors. Finally, Section 5.5 analyzes the results from the longitudinal deployment study and examines quotes from the caregivers about the CAST word scramble game.

5.1 Qualitative Assessment of CAST Usability

To decide if the claim of simplistic design and intuitive usage is correct, we examine the results of the qualitative survey (survey questions can be seen in Section 3.5). Recall that for questions 1 (responsiveness), 2 (intuitiveness), 3 (ease of use), and 8 (learnability), the

lowest score is a 1 and the highest score is an 5. The original claim stated in Section 3.1 was that the CAST app was simplistic in design and intuitive in usability. The two primary questions from the survey which address this claim are questions 2 (intuitiveness) and 3 (ease of use) which address the intuitiveness and design aspects, respectively. Table 4.1 shows the results of the survey and we note the mean scores for questions 2 (intuitiveness) and 3 (east of use) to be 4.47 and 4.47, respectively. These two averages which approach the maximum score of 5 support the claim of simplistic design and intuitive usage. Furthermore, both questions had a mode of 5 which shows that a majority of participants completely agreed with the claim. We also presented question 8 (learnability) which asked if the participants thought that most people would quickly and easily learn the CAST app. The resulting mean was 4.87 and mode was 5, which shows that our 15 participants thought that any individual is capable of operating the CAST word scramble game without trouble.

Five participants provided useful written feedback about the application design that was divided into three parts, which include the skip functionality, the app's keyboard, and the app's locked orientation. First, the skip function of the CAST word scramble game left three participants confused. Specifically, they wanted to return to the current word if they either accidentally skipped the word or decided they want to give it another attempt. Currently the user is not able to return to a word once the skip button has been pressed. Two of the participants provided this feedback as an additional comment in question 9 (additional suggestions) of the survey (quotes 5.1 and 5.2). The third participant answered "Yes" in question 4 (did not know how to proceed), which asked if the participant did not know how to proceed in the application, and left the corresponding comment (quote 5.3). This shows a possible limitation in the apps design. The first quote especially brings to attention an important aspect of "accidentally pressing skip." It is entirely possible, even probable, that when this app is given to an older adult that they will make mistakes during gameplay. This limitation though can be easily rectified in CAST version 2.0.

5.1. "No back button if accidentally press skip."

5.2. “When you push skip, make it able to go back to the question.”

5.3. “I didn’t know if I can skip and then return back.”

The fourth quote is provided by a participant which had a suggestion for the app’s keyboard functionality (quote 5.4). Specifically, when the user is finished typing in their answer for a words unscramble, they suggested that the keyboard disappear automatically. This is due to the confusion of having essentially two keyboard removal buttons, the “Done” key and the “Return” key which is represented by a left facing arrow. The participant clicked the “Return” key which did nothing as there is only one text field. Normally, the return key moves the cursor to the next text field for rapidly entering text into the app. The “Done” key is what closes the keyboard and allows for pressing the “Guess” or “Skip” buttons. The keyboard used in the CAST app is the standard Android keyboard which cannot be changed without the user downloading a new keyboard from the app store. Addressing this issue could be done by including a short and explicit instruction on the word scramble screen on how to close the keyboard to avoid confusion.

5.4. “When the user types his/her word and wants to return to the first (main) page of the app, they need to press ‘Done’ on the keyboard. But, it is easier if the keyboard disappears automatically to see the main page again.”

The last quote provides feedback on a quality of life design aspect. One participant left an additional comment in question 9 (additional suggestions) of the survey about rotating the orientation of the app (quote 5.5). Currently, the app is locked in landscape mode. When designing the CAST app, we thought it to be more user friendly to lock the app in landscape mode for two reasons. First, it prevents the app from accidentally rotating when slightly moving the device, causing an annoyance to the user. Next, it allows for a larger text font, larger buttons to be pressed, and a larger keyboard for entering answers into the word scramble game. We took into account older adults potential declining eyesight and

wanted to enlarge everything as much as possible while keeping a clean look to the user interface.

5.5. “Couldn’t rotate orientation.”

5.2 Analysis of Word Difficulty Levels

When choosing a word set for the CAST word scramble game, it is important that the words are discriminatory. That is, for any given word, the word is considered a discriminatory word if it is recognized as having the same difficulty level between a majority of users. A majority could be considered as approximately 70% or, 34/48 of users with respect to our dataset. These discriminatory words are important as they are used as indicators for an individual’s ability level and therefore, must belong to a distinct difficult. For example, if a user is given three words of Easy, Medium, and Hard discriminatory difficulty and they are only able to complete the Easy and Medium difficulties, it can deduced that this individual has a word scramble ability of Medium difficulty. From our chosen word set, we discuss three words which have this discriminatory quality (“water,” “check,” and “hazardous”) and one word which does not (“quakes”). Such discriminatory words are important as they act as predictors for other words in the same category. So, if a user were to unscramble two Medium words in a row, we can predict that if the user is given a third Medium word that he should be able unscramble it as well.

We begin with the words “water” and “check.” The solve rate for “water” is 32/47 with a mean URD of 4.0 and “check” with a solve rate of 42/47 and a mean URD of 2.6. With a clearly high unscramble rate for these two words and with mean URD’s belonging to the Easy category, as determined by the Rasch TM, we can conclude that both “water” and “check” are discriminatory words which belong in the Easy category. When using a combination of both the URD and RMD for deciding a words difficulty placement, each category can be filled accordingly. As shown in Table 4.2, there are multiple words belong-

ing to each category in which a subset have matching URD difficulty ratings. For example, the words “check,” “raspberry,” and “pistachio” have URD ratings of 41/47 for Easy, 34/47 for Medium, and 44/46 for Hard, respectively. These three URD ratings match the RMD difficulties as well with respect to the thresholds provided from the Rasch TM. This allows for a baseline of words to exist in the CAST word scramble game for the users to traverse through as their ability levels improve or decline. Though, a potential decline in word scramble performance of a user in a category could be used as an indicator for decline in overall task performance.

The word “hazardous” had two different scrambled configurations used in the CAST word scramble game. The first version had the ‘ous’ suffix unscrambled while v2 included the suffix in the scramble. As we anticipated that longer words would be more difficult, these two versions were included to see if the unscrambled suffix could lower the perceived difficulty of the word. For the “hazardous” v1, it was unscrambled by 39/48 participants while v2 was unscrambled by 19/48 participants. Not only was the “hazardous” word perceived as two different difficulties from the users perspective, the Rasch PM also came to the same conclusion. As seen in Table 4.2, the “hazardous” v1 has a RMD of 4 while v2 has a RMD of 6, placing them in the Easy and Hard categories, respectively. This shows that by leaving the suffix or potentially even the prefix unscrambled in a word, it allows for longer words to be included in the Easy or Medium categories. Furthermore, it shows that the same word scrambled in two different way can have separate difficulty ratings. This allows us to increase our word set by simply providing different scrambled versions of existing words rather than curating additional words.

Finally, during the selection process of the word set, we did not take into account that some scrambled words could unscramble to multiple words. The word “quakes” is the example of this with a solve rate of 24/48 and a mean URD of 7.3. This word is considered a bijective word, i.e. the letters could be unscrambled as two distinct words. This is an unintended outcome as each word scramble in the CAST game only has one intended unscram-

ble. When participants encountered this word, they were given the scramble "skuqea," for which multiple participants thought the unscrambled word was "squeak." This left multiple participants visually flustered during their play-through. This observation acts as a known limitation of the word set. Having a list of all possible word combinations for a given string of letters can rectify this issue for the future work. Finally, during gameplay, we noticed a common trend from our participants for when they encountered a word length of 7 letters or more, they exhibited visual frustration. From Table 4.2, we can see that many of the words in the hard category are generally longer such as "pistachio," "manatee," "gargoyle," etc. This observation also explains the negative skewness of 41/46 Hard ratings for the word "pistachio" as seen in Table 3.2.

5.3 Individualized Word Difficulty Predictions Using FIS

The performance presented in Section 4.3 is poorer than initially anticipated, though not entirely surprising from the initial observations of the participants playing the CAST word scramble game. Research by Linacre et al. showed that it is difficult to go from a dichotomous data system to a multi-categorical system such as three or more categories [25]. This is due to participants failing to respond to an instrument in a way expected by the system designers. In this thesis, this is represented by the overwhelming binary rating habits of the participants. Yang et al. performed research which utilized machine learning techniques to map objective health data to subjective pain scale ratings of sickle cell diseased patients [39]. The authors in that study encountered similar difficulties while predicting pain scores using the original 11 point pain scale (0-10), but improved the results when combining classes to a 4 point pain scale (no pain, low pain, moderate pain, and high pain) [39]. The studies by both Yang et al. [39] and Linacre et al. [25] show the difficulties in mapping objective data to subjective ratings in a non-binary classification system. However, our hypothesis requires a non-binary classification system to allow for improvement or decline in

the task performance with respect to the CAST word scramble game. A binary category system would make it more difficult to detect small, but significant, changes in the users performance over time, therefore at least a three-category system is desirable.

With the URD as the ground truth, the performance is overall poor as shown in Tables 4.3 and 4.4. Table 4.3 shows the Hard Precision of 0.94 or the Easy Recall of 0.95. We see the Medium category suffer in both resubstitution and leave-one-out cases. This is potentially due to the participants rating the words in unusual ways. That is, participants showed frustration if they were unable to unscramble a word in a short amount of time; this was reported after gameplay by multiple participants. This led the participants to rating the word in the higher range regardless if they eventually successfully unscrambled the word. These specific ratings do not only encapsulate the difficulty rating of a word, rather it includes the frustration the participant had during the task as well.

The second ground truth, the RMD, attempts to maneuver past this problem of binary and frustrated user ratings by taking into account only whether the participants correctly unscrambled the word or not. However, Tables 4.3 and 4.4 show that the performance of the RMD ground truth is poorer than that of the URD. We can see an F1 score of 0.45, 0.11, and 0.63 for the Easy, Medium, and Hard categories, respectively. Supervised machine learning techniques, such as the FIS, perform best with an equal distribution of class data. As such, the unequal distribution of words while using the RMD ground truth, as shown in Table 4.2, is the probable cause to the low performance. The Easy category contains 5 words while the Medium and Hard categories contain 7 and 16, respectively. Furthermore, as the FIS was heuristically created with the assumption that we would encounter an even distribution of difficulties, the binary split causes expected performance loss.

The previously discussed results of the URD and RMD ground truths in the resubstitution setting are similar to that of the leave-one-out setting. Of the 28 words in the word set, 11 were misclassified as false positives. From the false positives, 5 were misclassified as Easy from Medium and the last 6 were misclassified as Easy from Hard. These misclas-

sifications were caused by the FIS outputting an IWD of Easy. This reaffirms the problem of the even distribution assumption in conjunction with the unevenly distributed word categories. While the word categories are static as determined by the Rasch TM, we can adjust the membership functions of the heuristically created FIS using GA in an attempt to rectify the even distribution assumption.

5.4 Individualized Word Difficulty Predictions Using GA

We implemented the GA (discussed in Section 3.7) to improve the membership functions of the heuristically created FIS. Beginning with the URD ground truth, we examine Table 4.5 to see an F1 score improvement of +0.08 and +0.09 for the Medium and Hard resubstitution categories, respectively. The medium category for the URD improved slightly but is still low. When using a three category system, the middle category generally will perform the worst of the three [25]. We could convert the output of the FIS into a two category Easy and Medium system, though that would be detrimental to our hypothesis of progressing and regressing through *multiple* categories.

When examining the RMD ground truth in the resubstitution setting, a medium category arises as compared to the FIS without the GA implementation. As shown in Table 4.5, an improvement of +0.28, +0.37, and +0.36 for Precision, Recall, and F1 score is seen in the RMD resubstitution method over the heuristically created FIS results. Looking at the actual number of correct classifications out of the possible 1320, there are 145 correct Medium classifications as opposed to the 25 correct classifications of the heuristic FIS. While the Medium category improved, the Easy and Hard categories shifted in performance as well. The Easy category had a slight overall decline with a Precision, Recall, and F1 score change of +0.17, -0.32, and +0.05, respectively. Furthermore, the Hard category had an overall improvement with a Precision, Recall, and F1 score change of -0.02, +0.20, and +0.11, respectively. We can see that the performance improved dramatically for

Medium, and slightly for Hard and Easy using the F1 score metric which takes both precision and recall into account. These changes in performance for the RMD resubstitution method provide more consistency (i.e. a more prevalent Medium category in parallel with a well performing Easy and Hard category) across the three categories as well as an overall improvement in performance.

Both the URD and RMD ground truth labels provide different, but useful, information. As it is not feasible to educate an individual on making accurate ratings with respect to Medium and Hard words while preserving their subjectivity, the RMD shows promise. Conversely, when dealing with a distinctly Easy word, the URD is a prominent ground truth for the IWD output to compare against. Thus, a combination of both the URD and RMD can be used depending on an individuals baseline word scramble ability. For example, if a user is able effectively unscramble words in the Medium or Hard categories, then the RMD will be used. Conversely, if the user is able to unscramble words in only the Easy category, then the URD will be used.

5.5 Analysis of Longitudinal Caregiver Data

We examined the valuable temporal data about the CAST word scramble game that the dementia caregiver participants provided. Specifically, the results from the four Figures 4.3, 4.4, 4.5, and 4.6 show the IWD and URD for a subset of words from caregiver participants 1 and 2, respectively. Recall that the RMD is global for this specific word set, therefore it does not change and is not presented in the figures. These word subsets were chosen to signify the performance trend for both IWD and URD. Specifically, how both decline from week 1 to week 2. Beginning with the URD, we can see a more a more drastic shift of ratings from week 1 to week 2. Examples of this are from Figure 4.4 with the words "manatee" and "lavender" while Figures 4.6 has the words "khaki," "orange," "liberty," and "jasmine." As stated in Section 4.5, from the full word set of each participant, 15/28

and 8/24 words had URD ratings lower in week 2 from week 1 for participants 1 and 2, respectively. Of the remaining words which did not decrease, 11/28 and 11/24 URD stayed the same and 2/28 and 5/24 URD increased for participants 1 and 2, respectively. This trend of lower URD over time could suggest a familiarity with the words set as each participant saw the exact same order of words with the same scramble twice over a two week period.

Unlike the more significant differences seen in the URD over time, the IWD differentials from week 1 to week 2 are smaller in comparison. Even with the smaller decrease of IWD, a majority of the words did indeed decrease. From caregiver participant 1, 21/28 words decreased in IWD while 14/24 decreased for caregiver participant 2. The smaller IWD differences shown in Figures 4.3 and 4.5 from week 1 to week 2 can be explained as the participants had similar gameplay experiences each week. That is, the time taken and number of guesses is similar from week 1 to week 2. For example, the word "ebony" for participant 1 had 2 guesses and 41 seconds taken in week 1. Then, week 2 of "ebony" had 1 guess and 25 seconds. While the participant clearly showed an adaptation to that particular word and its corresponding scramble, the FIS did not see a drastic change from the two attempts. Also, the 2 participants only saw each word twice, which allows for some familiarity to be obtained but not a significant amount. As each game is spread out over a weeks time, the second time a word is encountered, there has been a seven day gap.

We hypothesize that if the participants continued the temporal deployment study for two more weeks, that the IWD would continue to drop from week to week as the caregivers gained more experience with the scrambles. These results and hypothesis are interesting as it shows a slow decline of IWD over time, indicating a learning of the scrambled word. The next iteration of CAST could include a mixture of pre-experienced words and new words. We would then expect a user to correctly unscramble the pre-experienced word/words in the gameplay session while giving more energy on a new unseen word. Failing to unscramble pre-experienced words could used as an indicator of declining task performance.

We also gathered quotes from discussing with the caregiver participants throughout

the longitudinal study. The first caregiver participant said that they became more familiar with the game and that it became easier over time (quote 5.6). Participant one also said that they loved that the words repeated from week 1 to week 2 of gameplay (quote 5.7). Furthermore, caregiver participant 2 mentioned that they researched strategies on how to complete word scrambles before beginning the game (quote 5.8). They also entered the scrambles into a website to figure out what the answer could be after finishing a gameplay session (quote 5.8). Quotes 5.6 and 5.8 from caregiver participants 1 and 2, respectively, add to the observation that participants gained familiarity with the words from week 1 to week 2. This supports the reasoning for the general decrease in IWD as shown in Figures 4.3 and 4.5. Finally, Quote 5.7 from caregiver participant 1 supports the idea of including a mixture of pre-experienced words and new words in a day of gameplay.

5.6. "I became familiar with it and it wasn't as umm it seemed easier."

5.7. "I loved that the words repeated."

5.8. "Before I started I went online in order to get strategies to do word scrambles so that has helped with some of the words and then, I don't cheat, but afterward I go in and type them in and find out what it was."

6

Conclusion

In this thesis, we created the CAST word scramble game as well as a FIS to output the IWD of a word. The FIS was heuristically constructed under the assumption of having an even distribution of participant abilities and word difficulties. The FIS acted as a way to assign an IWD to a given word with its corresponding scramble and to also place words into distinct difficulty categories. We utilized two separate ground truths to compare our results which are the URD and RMD. Promising results were obtained when comparing the FIS IWD output to both the URD and RMD ground truths. We examined the URD which showed a split bias rating from our participant cohort (n=48). This observation went against our assumptions made for the FIS and resulted in poorer performance using the standard performance metrics of Precision, Recall, and F1 score. To rectify this issue, we utilized the Rasch PM to calculate the RMD ground truth. Comparing against the RMD, there were minor changes with respect to the original FIS. Finally, we implemented GA in order to update the membership functions and provide better system performance. Comparing the URD and RMD to the GA improved FIS, the results improved greatly.

We have shown that with the GA improvements to the FIS, it is indeed possible to provide an individualized experience for any given user. Furthermore, by utilizing both the URD and RMD ground truths in conjunction, this allows for tracking changes in perfor-

mance via changes in the users gameplay experience. Research has shown that mapping objective data, in our case the data collected via the CAST word scramble game, to subjective ratings is a difficult task.

In this IRB approved study, three research questions were posed and answered throughout this thesis. First, as discussed in Section 3.6, an FIS was constructed with the ability of measuring the IWD of a given word and its corresponding scramble (RQ1). Second, Section 5.2 shows that discriminatory words such as "check," "raspberry," and "pistachio" exist for the Easy, Medium, and Hard difficulties, respectively (RQ2). Finally, Sections 4.3 and 4.4 show adequate results for the FIS IWD output compared to both the URD and RMD ground truths (RQ3). With this information, the CAST application can allow for unobtrusive temporal monitoring of task performance. Specifically, it is to be focused on the dementia caregiver for continuous task performance assessment. Thus, allowing early community intervention to improve both the caregiver and patient outcomes.

7

Future Work

As this thesis focuses on the feasibility of implementing an unobtrusive approach to monitoring task performance, the planned future work is extensive.

1. Increase the size of our participant cohort.

2. Deploy CAST to additional dementia caregivers on a longer time scale.

3. Add more words to the word set.

4. Measure the cortisol level of participants before and after gameplay.

Beginning with the first item, we would first like to increase the size of our cohort to further generalize the systems performance. With a larger sample size, it is possible that the negative observations we observed could average out or slightly diminish. Following, the deployment of the CAST app allows for soliciting feedback from the target population as well as allowing for gathering data on the application usage. Next, we would like to increase the overall size of the word set, such as to 500 or more words. During this process, we would also like to account for multiple target unscrambles to avoid the "quakes" and "squeak" issues discussed in Section 5.2. Also, drastically increasing the word set would help prevent the user from learning the words and their scrambles. Finally, due to IRB

constraints, measuring the cortisol levels of participants was neither feasible nor possible. Thus, we would like to measure the cortisol levels before and after gameplay to see the corresponding stress of a participant during gameplay.

Bibliography

[1] Alzheimer's Association. 2018 Alzheimer's disease facts and figures. https://www.alz.org/facts/overview.asp

[2] Alzheimer's Association. Types of Dementia. https://www.alz.org/dementia/types-of-dementia.asp

[3] Tanvi Banerjee, Jennifer C Hughes, Larry Lawhorne, Amit Sheth, Matthew Peterson, and Krishnaprasad Thirunarayan. Dementia. http://wiki.knoesis.org/index.php/Dementia

[4] William J Boone. Rasch analysis for instrument development: Why, when, and how? *CBE-Life Sciences Education*, 15(4):rm4, 2016.

[5] Henry Brodaty, Charles McGilchrist, Lynne Harris, and Karin E Peters. Time until institutionalization and death in patients with dementia: role of caregiver training and risk factors. *Archives of Neurology*, 50(6):643–650, 1993.

[6] Alistair Burns and Steve Iliffe. Alzheimer's disease. *BMJ*, 338, 2009.

[7] AA Burstein, O DaDalt, B Kramer, LA D'Ambrosio, and JF Coughlin. Dementia caregivers and technology acceptance: Interest outstrips awareness. *Gerontechnology*, 14(1):45–56, 2015.

[8] Connie Canam and Sonia Acorn. Quality of life for family caregivers of people with chronic health problems. *Rehabilitation Nursing*, 24(5):192–200, 1999.

[9] Sheldon Cohen, T Kamarck, R Mermelstein, et al. Perceived stress scale. *Measuring stress: A Guide for Health and Social Scientists*, 1994.

[10] Alison Dillon, Mark Kelly, Ian H Robertson, and Deirdre A Robertson. Smartphone applications utilizing biofeedback can aid stress reduction. *Frontiers in Psychology*, 7:832, 2016.

[11] G Downing. Biomarkers definitions working group. biomarkers and surrogate endpoints. *Clinical Pharmacology & Therapeutics*, 69:89–95, 2001.

[12] Howard Feldman, Serge Gauthier, Jane Hecker, Bruno Vellas, Birol Emir, Vera Mastey, and Ponni Subbiah. Efficacy of donepezil on maintenance of activities of daily living in patients with moderate to severe alzheimer's disease and the effect on caregiver burden. *Journal of the American Geriatrics Society*, 51(6):737–744, 2003.

[13] Janna M Glozman. Quality of life of caregivers. *Neuropsychology Review*, 14(4):183–196, 2004.

[14] Michel Goedert and Maria Grazia Spillantini. A century of alzheimer's disease. *Science*, 314(5800):777–781, 2006.

[15] Charles L Gutshall, David P Hampton Jr, Ismail M Sebetan, Paul C Stein, and Thomas J Broxtermann. The effects of occupational stress on cognitive performance in police officers. *Police Practice and Research*, 18(5):463–477, 2017.

[16] Richard W Hamming. Error detecting and error correcting codes. *Bell Labs Technical Journal*, 29(2):147–160, 1950.

[17] Erik Harpstead, Christopher J MacLellan, Kenneth R Koedinger, Vincent Aleven, Steven P Dow, and Brad A Myers. Investigating the solution space of an open-ended

educational game using conceptual feature extraction. In *Educational Data Mining 2013*, 2013.

[18] Erik Harpstead, Brad A Myers, and Vincent Aleven. In search of learning: facilitating data analysis in educational games. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 79–88. ACM, 2013.

[19] Christoffer Holmgard, Georgios N Yannakakis, Karen-Inge Karstoft, and Henrik Steen Andersen. Stress detection for ptsd via the startlemart game. In *Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on*, pages 523–528. IEEE, 2013.

[20] Jennifer C Hughes, Tanvi Banerjee, Garrett Goodman, and Larry Lawhorne. A preliminary qualitative analysis on the feasibility of using gaming technology in caregiver assessment. *Journal of Technology in Human Services*, 35(3):183–198, 2017.

[21] Mahdi Jampour, Mehdi Ziari, Reza Ebrahim Zadeh, and Maryam Ashourzadeh. Impulse noise detection and reduction using fuzzy logic and median heuristic filter. In *Networking and Information Technology (ICNIT), 2010 International Conference on*, pages 19–23. IEEE, 2010.

[22] Jason Kane. Health costs: How the u.s. compares with other countries.

[23] Nicole Korten, Hannie C Comijs, Brenda WJH Penninx, and Dorly JH Deeg. Perceived stress and cognitive function in older adults: which aspect of perceived stress is important? *International Journal of Geriatric Psychiatry*, 32(4):439–445, 2017.

[24] Ari Levine, Orna Zagoory-Sharon, Ruth Feldman, John G Lewis, and Aron Weller. Measuring cortisol in human psychobiological studies. *Physiology & Behavior*, 90(1):43–53, 2007.

[25] John M Linacre et al. Optimizing rating scale category effectiveness. *J Appl Meas*, 3(1):85–106, 2002.

[26] Yueh-Feng Yvonne Lu and May Wykle. Relationships between caregiver stress and self-care behaviors in response to symptoms. *Clinical Nursing Research*, 16(1):29–43, 2007.

[27] Neil Maiden, Sonali D'Souza, Sara Jones, Lars Müller, Lucia Pannese, Kristine Pitts, Michael Prilla, Kevin Pudney, Malcolm Rose, Ian Turner, et al. Computing technologies for reflective, creative care of people with dementia. *Communications of the ACM*, 56(11):60–67, 2013.

[28] Jeff Makowka, Theodora Lau, Stan Kachnowski, Laura Pugliese, Molly Woodriff, Margaret Griffin, Olga Crowley, Vivian Lam, Gabriel Lee, and Nat Harward. Caregivers & technology: What they want and need. *AARP Real Possibilities Project CATALYST the Power of We*, pages 1–42, 2016.

[29] Srinivasa Narayan and Yuri Owechko. Fuzzy expert system for interpretable rule extraction from neural networks, May 13 2003. US Patent 6,564,198.

[30] TRULS ØSTBYE, SUZANNE TYAS, IAN MCDOWELL, and JOHN KOVAL. Reported activities of daily living: agreement between elderly subjects with and without dementia and their caregivers. *Age and Ageing*, 26(2):99–106, 1997.

[31] Hedieh Ranjbartabar, Amir Maddah, and Manolya Kavakli. emedicalhelp: A customized medical diagnostic application: Is a single questionnaire enough to measure stress? In *Computer and Information Science (ICIS), 2015 IEEE/ACIS 14th International Conference on*, pages 367–372. IEEE, 2015.

[32] Samsung. Galaxy tab a 10.1.

[33] Finnegan Southey, Gang Xiao, Robert C Holte, Mark Trommelen, and John W Buchanan. Semi-automated gameplay analysis by machine learning. In *AIIDE*, pages 123–128, 2005.

[34] David A Squires. The us health system in perspective: a comparison of twelve industrialized nations. *Issue Brief (Commonwealth Fund)*, 16:1–14, 2011.

[35] Kyle Strimbu and Jorge A Tavel. What are biomarkers? *Current Opinion in HIV and AIDS*, 5(6):463, 2010.

[36] Frank M Torti Jr, Lisa P Gwyther, Shelby D Reed, Joëlle Y Friedman, and Kevin A Schulman. A multinational review of recent trends and reports in dementia caregiver burden. *Alzheimer Disease & Associated Disorders*, 18(2):99–109, 2004.

[37] Minh Khue Phan Tran, Philippe Robert, and François Bremond. A virtual agent for enhancing performance and engagement of older people with dementia in serious games. In *Workshop Artificial Compagnon-Affect-Interaction 2016*, 2016.

[38] Annibal Truzzi, Letice Valente, Ingun Ulstein, Eliasz Engelhardt, Jerson Laks, and Knut Engedal. Burnout in familial caregivers of patients with dementia. *Revista Brasileira de Psiquiatria*, 34(4):405–412, 2012.

[39] Fan Yang, Tanvi Banerjee, Kalindi Narine, and Nirmish Shah. Improving pain management in patients with sickle cell disease from physiological measures using machine learning techniques. *Smart Health*, 2018.

[40] Lotfi A Zadeh. Outline of a new approach to the analysis of complex systems and decision processes. *IEEE Transactions on Systems, Man, and Cybernetics*, (1):28–44, 1973.